Technical Report 1198

Army Enlisted Personnel Competency Assessment Program: Phase III Pilot Tests

Karen O. Moriarty and Deirdre J. Knapp (Editors)
Human Resources Research Organization

March 2007

United States Army Research Institute for the Behavioral and Social Sciences

U.S. Army Research Institute
for the Behavioral and Social Sciences

A Directorate of the Department of the Army
Deputy Chief of Staff, G1

Authorized and approved for distribution:

SCOTT E. GRAHAM
Acting Technical Director

MICHELLE SAMS
Director

Research accomplished under contract
for the Department of the Army

Human Resources Research Organization

Technical review by

Rachel Mapes, U.S. Army Research Institute
Stephanie T. Muraca, U.S. Army Research Institute

NOTICES

DISTRIBUTION: Primary distribution of this Technical Report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, Attn: DAPE-ARI-MS, 2511 Jefferson Davis Highway, Arlington, Virginia 22202-3926.

FINAL DISPOSITION: This Technical Report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this Technical Report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

1. REPORT DATE (dd-mm-yy)	2. REPORT TYPE	3. DATES COVERED (from. . . to)
March 2007	Interim	January 2005 – January 2006

4. TITLE AND SUBTITLE	5a. CONTRACT OR GRANT NUMBER
Army Enlisted Personnel Competency Assessment Program: Phase III Pilot Tests	DASW01-03-D-0015/DO 0013
	5b. PROGRAM ELEMENT NUMBER 622785
6. AUTHOR(S)	5c. PROJECT NUMBER A790
Karen O. Moriarty and Deirdre J. Knapp (Editors) (Human Resources Research Organization)	5d. TASK NUMBER 104
	5e. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
Human Resources Research Organization 66 Canal Center Plaza, Suite 400 Alexandria, VA 22314	IR-05-74
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. MONITOR ACRONYM
U. S. Army Research Institute for the Behavioral & Social Sciences 2511 Jefferson Davis Highway Arlington, VA 22202-3926	ARI
	11. MONITOR REPORT NUMBER Technical Report 1198

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
Contracting Officer's Representatives and Subject Matter POCs: Tonia Heffner and Peter Greenston
Contract for Manpower, Personnel, Leader Development, and Training (COMPLETRS) for the U.S. Army Research Institute.

14. ABSTRACT *(Maximum 200 words)*:

In the early 1990s, the Department of the Army abandoned its Skill Qualification Test (SQT) program due primarily to maintenance, development, and administration costs. This left a void in the Army's capabilities for assessing job performance qualification. To meet this need, the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) instituted a 3-year program of feasibility research related to the development of a Soldier assessment system that is both effective and affordable. The PerformM21 program has two mutually supporting tracks. The first focuses on the design of a testing program and identification of issues related to its implementation. The second track is a demonstration of concept – starting with a prototype core assessment targeted to all Soldiers eligible for promotion to Sergeant, followed by job-specific prototype assessments for several Military Occupational Specialties (MOS). The prototype assessments were developed during the first 2 years of the research program. The present report describes work conducted in the final year of the PerformM21 program, in which five prototype MOS-specific assessments (along with the common core examination) were pilot tested on a sample of specialists/corporals.

15. SUBJECT TERMS

Behavioral and social science	Personnel	Job performance measurement
Manpower	Competency assessment	

SECURITY CLASSIFICATION OF			19. LIMITATION OF ABSTRACT	20. NUMBER OF PAGES	21. RESPONSIBLE PERSON
16. REPORT	17. ABSTRACT	18. THIS PAGE			
Unclassified	Unclassified	Unclassified	Unlimited	66	Ellen Kinzer Technical Publications Specialist (703) 602-8047

Standard Form 298

Technical Report 1198

Army Enlisted Personnel Competency Assessment Program: Phase III Pilot Tests

Karen O. Moriarty and Deirdre J. Knapp (Editors)
Human Resources Research Organization

Selection and Assignment Research Unit
Michael G. Rumsey, Chief

U.S. Army Research Institute for the Behavioral and Social Sciences
2511 Jefferson Davis Highway, Arlington, Virginia 22202-3926

March 2007

Army Project Number
622785A790

Personnel Performance
and Training Technology

Acknowledgements

U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) Contracting Officer Representatives (COR)

Dr. Tonia Heffner and Dr. Peter Greenston of ARI served as co-COR for this project, but their involvement and participation went far beyond the usual COR requirements. Their contributions and active input played a significant role in the production of the final product and they share credit for the outcome. Of particular note are their activities in conveying information about the project in briefings and presentations to Army Leadership on many important levels.

The Army Test Program Advisory Team (ATPAT)

The functions and contributions of the ATPAT, as a group, are documented in this report. But this does not fully reflect the individual efforts that were put forth by members of this group. Project staff is particularly indebted to Sergeant Major Michael Lamb, currently with Army G-3, who served as the ATPAT Chairperson during this work.

The other individual members of the ATPAT who were active and involved during this phase were:

SGM Ron Pruyt, Co-Chair
SGM Bill Bissonette
SGM John Cross
SGM Osvaldo Del Hoyo
SGM Daniel Dupont
SGM (R) Julian Edmondson
CSM Dan Elder
CSM Mark Farley
CSM Gary Ginsburg
SGM John Griffin
SGM John Heinrichs
SGM (R) James Herrell
SGM Enrique Hoyos
CSM Nick Piacentini
SGM Tony McGee
SGM David Litteral
SGM John Mayo
SGM Pamela Neal
CSM Rock Orozco
SGM Tim Ozman
CSM Doug Piltz
SGM (R) Gerald Purcell
CSM Robie Roberson
SGM Terry Sato
CSM Otis Smith Jr
SGM Irene Torkildson
MSG Edward Herbert
MSG Matthew Northen
SFC Kevin Barney
SFC Martha Chavez
Mr. Jeff Colimon

ARMY ENLISTED PERSONNEL COMPETENCY ASSESSMENT PROGRAM:
PHASE III PILOT TESTS

EXECUTIVE SUMMARY

Research Requirement:

The Army Training and Leader Development Panel Non-Commissioned Officer (NCO) survey (Department of the Army, 2002) called for objective performance assessment and self-assessment of Soldier technical and leadership skills to meet emerging and divergent Future Force requirements. The Department of the Army's previous experiences with job skill assessments in the form of Skill Qualification Tests (SQT) and Skill Development Tests (SDT) were reasonably effective from a measurement aspect but were burdened with excessive manpower and financial resource requirements.

Procedure:

The U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) conducted a 3-year feasibility effort to identify viable approaches for the development of a useful yet affordable operational performance assessment system for Army enlisted personnel. Such a system would depend on technological advances in analysis, test development, and test administration that were unavailable in the previous SQT/SDT incarnations.

ARI's *Performance Measures for the 21st Century* research project (known as PerformM21) entailed three phases:

- Phase I: Identify User Requirements, Feasibility Issues, and Alternative Designs
- Phase II: Develop and Pilot Test Prototype Measures
- Phase III: Evaluate Performance Measures, Conduct a Cost-Benefit Analysis, and Make System Recommendations

The objective of Phase I was to identify issues that the overall recommendation needs to take into account for a viable, Army-wide system (Knapp & Campbell, 2004). Phase I also produced a rapid prototype assessment covering Army-wide "core content" with associated test delivery and test preparation materials (R. C. Campbell, Keenan, Moriarty, Knapp, & Heffner, 2004).

In Phase II, the research team (a) pilot tested the core competency assessment, (b) developed competency assessment prototypes for five Military Occupational Specialties (MOS), and (c) explored issues further to develop more detailed recommendations related to the design and feasibility of a new Army enlisted personnel competency assessment program. The work in Phase II is documented in Knapp and Campbell (2006).

In Phase III, the MOS tests (along with a short version of the common core examination) were pilot tested and a cost and benefit analysis of a notional Army program was conducted. The present report documents the pilot test activities.

Findings:

The prototype MOS assessments were successfully administered to approximately 500 E4 Soldiers in five MOS: Patriot Air Defense Control Operators/Maintainers (14E), Armor Crewman (19K), Military Police (31B), Wheeled Vehicle Mechanic (63B), and Health Care Specialist (91W). These assessments included job knowledge tests enhanced with advanced graphics features, situational judgment tests, and simulations. We also administered a short version of the common core examination to Soldiers in the five target MOS plus 244 Soldiers in other MOS. Except for the 14E simulation, the tests were web-based and delivered primarily through Army Digital Training Facilities. Our experience with the different test methods was consistent with our prior expectations about their respective strengths and weaknesses. For example, the job knowledge tests provided a relatively inexpensive strategy for broadly covering job requirements whereas the computer-based simulation for Patriot Air Defense Control Operators/Maintainers provided a more realistic work sample that was enthusiastically received by Soldiers, but at greater cost and with considerably less comprehensive coverage of job requirements. Cost considerations aside, use of multiple measurement methods in an MOS would be a desirable option.

Utilization and Dissemination of Findings:

The assessment work has resulted in lessons learned and test item banks suitable for incorporation into an operational test program. The lessons learned include all portions of an operational test program, from Soldier notification to providing Soldier feedback. The program design and technology issues and recommendations are intended to help Army leaders make informed decisions about an operational competency assessment program.

ARMY ENLISTED PERSONNEL COMPETENCY ASSESSMENT PROGRAM:
PHASE III PILOT TESTS

CONTENTS

List of Tables

List of Figures

ARMY ENLISTED PERSONNEL COMPETENCY ASSESSMENT PROGRAM: PHASE III PILOT TESTS

CHAPTER 1: PERFORMM21 RESEARCH PROGRAM OVERVIEW

Deirdre J. Knapp and Roy C. Campbell

Introduction

Individual Soldier readiness is the foundation of a successful force. In the interest of promoting individual Soldier performance, the U.S. Department of the Army has previously implemented assessment programs to measure Soldier knowledge and skill. The last incarnation of such a program was the Skill Qualification Test (SQT) program. The SQT program devolved over a number of years, however, and in the early 1990s the Army abandoned it due primarily to maintenance, development, and administration costs.

Cancellation of the SQT program left a void in the Army's capabilities for assessing job performance qualification. This was illustrated most prominently in June 2000, when the Chief of Staff of the Army established the Army Training and Leader Development Panel (ATLDP) to chart the future needs and requirements of the Noncommissioned Officer (NCO) corps. After a 2-year study, which incorporated the input of 35,000 NCOs and leaders, a major conclusion and recommendation was that the Army should: "Develop and sustain a competency assessment program for evaluating Soldiers' technical and tactical proficiency in the military occupational specialty (MOS) and leadership skills for their rank" (Department of the Army, 2002).

The impetus to include individual Soldier assessment research in the U.S. Army Research Institute for the Behavioral and Social Sciences's (ARI's) programmed requirements began prior to 2000, and was based on a number of considerations regarding requirements in Soldier selection, classification, and qualifications. For example, lack of operational criterion measures has limited improvements in selection and classification systems. Meanwhile, several significant events within the Army reinforced the need for efforts in this area. As a result of the aforementioned ATLDP recommendation, the Office of the Sergeant Major of the Army (SMA) and the U.S. Army Training and Doctrine Command (TRADOC) initiated a series of reviews and consensus meetings with the purpose of instituting a Soldier competency assessment test. Ongoing efforts within the Army G1 to revise the semi-centralized promotion system (which promotes Soldiers to the grades of E5 and E6) also investigated the feasibility of using performance (test)-based measures to supplement the administrative criteria that determine promotion. Ultimately, the three interests (ARI, SMA/TRADOC, G1) coalesced; the ARI project sought to incorporate the program goals and operational concerns of all of the Army stakeholders while still operating within its research-mandated orientation.

To meet the Army's need for job-based performance measures and identify viable approaches for the development of an effective and affordable Soldier assessment system, ARI instituted a 3-year program of feasibility research called *Performance Measures for the 21st Century* (PerformM21). This research was conducted with contract support from the Human

Resources Research Organization (HumRRO) and its subcontractors, Job Performance Systems, Inc, The Lewin Group, and the SAG Corporation.

Research Program Overview

The PerformM21 research program is best viewed as having two mutually supporting tracks. The first track involved the conceptualization and capture of issues, features, and capabilities in Army testing design recommendations. The second track led to the development and administration of prototype tests and associated materials. These prototypes include both an Army-wide "common core" assessment and selected MOS tests. They are intended to reflect, inasmuch as possible, design recommendations for the future operational assessment program. Experiences with the prototypes, in turn, influenced elaboration and modification of the operational program design recommendations as they developed during the course of the 3-year research program.

Formally, PerformM21 had three phases:

- Phase I: Identify User Requirements, Feasibility Issues, and Alternative Designs
- Phase II: Develop and Pilot Test Prototype Measures
- Phase III: Evaluate Performance Measures, Conduct a Cost-Benefit Analysis, and Make System Recommendations

Phase I of PerformM21 resulted in program design recommendations that included such considerations as how an Army assessment would be delivered, how assessments would be designed, developed, and maintained, and what type of feedback would be given. We also developed a demonstration common core assessment test to serve as a prototype for the envisioned new Army testing program. This core assessment is a computer-based, objective test that covers core knowledge areas applicable to Soldiers in all MOS (training, leadership, common tasks, history/values). Phase I was completed in January 2004, and is documented in two ARI publications (R. C. Campbell, Keenan, Moriarty, Knapp, & Heffner, 2004; Knapp & Campbell, 2004).

Phase II of the PerformM21 program (which corresponds roughly to year two of the 3-year overall effort) had three primary goals:

- Conduct an operational pilot test of the common core assessment with approximately 600 Soldiers.
- Investigate job-specific competency assessments. This resulted in prototype assessments for five MOS.
- Continue to refine and to develop discussion and recommendations related to the design and feasibility issues established in Phase I.

The Phase II work is detailed in an ARI technical report edited by Knapp and Campbell (2006). Development of the MOS-specific prototype assessments is summarized in Chapter 2 of the present report.

The primary activities in Phase III were to pilot test the prototype MOS-specific assessments (as well as further pilot testing of the common core test) and to conduct a cost-benefit analysis of the notional assessment program. The pilot test work is detailed in the present report.

Related Efforts

In addition to the core elements of PerformM21 broadly outlined in the three phases, two related studies were generated by requirements uncovered during the PerformM21 research. The first was an analysis to determine the best way to help Soldiers gauge their overall readiness for promotion, including identification of strengths and weaknesses prior to testing (Keenan & Campbell, 2005). This research produced a prototype self-assessment tool intended to help prepare Soldiers for subsequent assessment on the common core test.

The second was an analysis designed to determine new or refocused skills and tasks associated with operations in Iraq and Afghanistan and to incorporate those in a common core assessment program. This study produced two major products. One was a prototype field survey designed to support the development of a common core test "blueprint," and the second was the development of additional common core test items targeted to content areas suggested by lessons learned in recent deployment operations. This work is documented in Moriarty, Knapp, and Campbell (2006).

The Army Test Program Advisory Team (ATPAT)

Early in Phase I, ARI constituted a group to advise us on the operational implications of Army assessment testing, primarily as part of the needs analysis aspect of the project. This group is called the Army Test Program Advisory Team (ATPAT), and is comprised primarily of Command Sergeants Major and Sergeants Major. ATPAT members represent key constituents of various Army commands and all components. After the needs analysis, the ATPAT assumed the role of oversight group for the common core and MOS assessments, and served as a resource for identifying and developing content for the tests. Eventually, the group became an all-around resource for all matters related to potential Army testing. The ATPAT also served as a conduit to explain and promote the PerformM21 project to various Army agencies and constituencies.

Research Approach: Integrating Process and Results

To structure the needs analysis process, project staff drafted a list of requirements for supporting an assessment program. Figure 1.1 lists the key components of the organizing structure, which is more fully explained in the Phase I needs analysis report (Knapp & Campbell, 2004). This structure helped organize our thinking and suggested the questions we posed to those providing input into the process. We obtained input from many sources as we considered the issues, ideas, and constraints associated with each requirement listed in Table 1.1. Thus, this needs analysis organizing structure was used as a foundation for conceptualizing details of a notional Army test system.

Table 1.1. Outline of PerformM21 needs analysis organizing structure

- Purpose/goals of the testing program
- Test content
- Test design
- Test development
- Test administration
- Interfacing with candidates
- Associated policies
- Links to Army systems
- Self-assessment

Our experience designing and developing prototype assessments informed our program design recommendations and associated cost estimates. A prime example of how this approach worked is illustrated by the development of the prototype MOS tests discussed in the present report. We deliberately tried-out different methods and different tactics to see what would work and what would not, knowing that the process would appear somewhat chaotic while it was ongoing, but that the approach could uncover novel procedures and results. When viewed independently of one another, the approach to each discrete MOS looks like just bits and pieces. However, when all experiences are put together *post facto,* they form a more coherent whole. Such was the case with the entire project.

Overview of Report

Chapter 2 summarizes the Phase II test development activities that generated the MOS-specific assessments pilot tested in Phase III. The remainder of the report describes the process and results of the Phase III prototype assessment pilot tests. Chapter 3 describes the data collection process and procedures. Chapters 4 through 6 discuss results for each of the major measurement methods used (i.e., job knowledge tests, situational judgment tests, and simulations). Chapter 7 looks at relations among scores yielded by the different test methods. Finally, Chapter 8 closes with an overall summary and discussion of results.

CHAPTER 2: OVERVIEW OF PERFORMM21 MOS TEST DEVELOPMENT PROCESS

Karen O. Moriarty and Deirdre J. Knapp

Introduction

The goal of the MOS-specific portion of the PerformM21 project was to explore the potential of different testing methods for job-specific testing for five different MOS. We restricted the development effort to selected prototype tests for each MOS and did not attempt to create tests that comprehensively covered each MOS's job requirements. The test audience in our research was Soldiers eligible for promotion to sergeant (E5), which we operationalized as E4 Soldiers with approximately 3 years time in service. The remainder of this chapter summarizes the test development efforts from Phase II.

It was our goal to select MOS for which diverse prototype assessment items could be created. To this end, we relied on two sources: (a) the work of Rosenthal, Sager, and Knapp (2005), which identified groupings of MOS based on the most effective assessment methods for each group and (b) the guidance of the ATPAT. Table 2.1 shows the MOS selected because of the opportunities and challenges they presented. The 14E (Patriot Air Defense Control Operator/Maintainer) and 19K (Armor Crewmen) MOS each offered the potential for high fidelity simulation test development. The 31B (Military Police) MOS presented a challenge because there are two distinct types of assignments in which Soldiers can find themselves. One is law enforcement and the other is peacekeeping/combat support, and it is possible that these two assignments require different sets of competencies[1]. Both 63B (Wheeled Vehicle Mechanic) and 91W (Health Care Specialist) have related civilian credentialing programs and were undergoing consolidation. Additionally, they were each ideal for trying principle or systems-based testing, which is a departure from the Army's standard, task-based testing.

Table 2.1. PerformM21 Target MOS

MOS	Proponent Location
14E Patriot Air Defense Control Operator/Maintainer	Fort Bliss, TX
19K Armor Crewman	Fort Knox, KY
31B Military Police	Fort Leonard Wood, MO
63B Wheeled Vehicle Mechanic	Aberdeen Proving Ground, MD
91W Health Care Specialist	Fort Sam Houston, TX

Job Analysis and Test Design

In order to develop test methods appropriate for each discrete MOS, particular attention was paid to the job analysis process. To identify the measurement methods most appropriate for each MOS, we implemented Rosenthal et al.'s (2005) method of conducting an inexpensive, highly standardized, preliminary job analysis. Under operational conditions, a comprehensive, full-scale job analysis follows the preliminary investigation. However, in this research, we chose to focus our resources on item development rather than conducting a full-scale job analysis for each MOS. Thus, we supplemented the Rosenthal et al. method with more focused information

[1] As it turns out, we were also able to adapt a training simulator to test some important 31B tasks.

obtained from a small number of subject matter experts (SMEs) to support development of the selected prototype measures.

We implemented Rosenthal's method using four sets of generic job descriptors. For the most part, these descriptors were based on the taxonomies that are part of the Occupational Network (O*NET) database maintained by the U.S. Department of Labor (Peterson, Mumford, Borman, & Fleischman, 1999). Following are descriptions of each descriptor set:

- Work context descriptors – examples include level of social interaction, attention to detail, or time pressure/decision speed.
- Cognitive complexity indicators – judgment/problem-solving, information intensity, and systems thinking.
- Knowledge requirements -
    - Declarative knowledge – knowledge of facts and things (i.e., knowing *what* to do).
    - Procedural knowledge – knowledge or skill at performing physical or psychomotor tasks (i.e., knowing *how* to do it).
- Generalized work activities – set of 34 abstract task-like statements (e.g., "monitoring processes, materials, or surroundings" and "estimating the quantifiable characteristics of products, events, or information").

In addition to having SMEs provide ratings for these job descriptors, we asked them to generate examples of particularly effective or ineffective performance (i.e., critical incidents) and to review and revise MOS-specific task lists that we created. We had developed the task lists by reviewing MOS references such as Soldier training publications (STP), field manuals (FMs), technical manuals (TMs), and other appropriate specific references. The SME-generated critical incidents and the revised task lists were used to develop test content and/or to help SMEs think comprehensively about their MOS requirements. Our efforts to encourage SMEs to be discriminating in their ratings were not successful as shown by the relatively high ratings of most of the descriptors. However, while these ratings were not very helpful in making decisions about test methods, they were valuable in getting SMEs to think comprehensively about job requirements. A full discussion of the process followed is available in the Phase II report (Knapp & Campbell, 2006).

Full-Scale Job Analysis

We had two opportunities to conduct a fuller scale job analysis using a survey approach. In Phase II, we developed and administered a web-based job analysis survey for the 31B MOS. The primary objective of this analysis was to investigate how the Army's training-oriented occupational analysis process (the Occupational Data Analysis, Requirements, and Structure [ODARS] program) could be adapted to provide data for developing test specifications. Of particular interest was using the survey results to develop a prototype blueprint for a test to evaluate the competence of E4 31B Soldiers eligible for promotion to the E5 pay grade.

Complete data were collected from 386 31B supervisors (E5/E6 pay grades) and 44 incumbents (E4 pay grade). Analysis of the survey data revealed that the tasks varied greatly in their importance to performance as an E4 MP Soldier, and that different groups of survey

respondents (e.g., supervisors and incumbents with and without recent deployment experience), appeared to largely agree with each other about the relative importance of the tasks to E4 pay grade job performance. The survey results were used to design a prototype blueprint that specifies the percentage of test content (for an E4 pay grade competency assessment) that should be devoted to each task category. The Phase II report (Knapp & Campbell, 2006) provides a fuller description of the 31B survey's development and administration. A detailed report of the findings was provided to the 31B proponent. The executive summary from the proponent report is provided here as Appendix A.

We had another opportunity to develop and administer a prototype test design survey, this time for Army-wide test content, as part of a related project. The approach, findings, and associated recommendations are provided in Moriarty et al. (2005).

Identification of Test Methods

In the process of identifying test methods, we focused on developing and evaluating test methods across the diverse MOS rather than developing all of the appropriate measures suggested by Rosenthal et al. (2005). Decisions about which measurement methods to try were ultimately made by the proponent SMEs and POCs with guidance from testing professionals in a process that we expect would mirror what would occur in an operational application. The test methods selected as a result of this decision-making process applied to each of the target MOS are shown in the matrix in Table 2.2.

Table 2.2. Assessment Methods by MOS

Method	14E	19K	31B	63B	91W
Expert evaluation of actual work products			(X)[a]		
Hands-on work sample tests		(X)[b]			
Computer-based simulations	X	X	X		
Multiple-choice simulations					X
Situational judgment tests			X		X
Multiple-choice tests (incorporating visual aids and audio/video clips and non-traditional item formats such as matching, ranking, and drag-and-drip)		X		X	X

[a]We explored the possibility of scoring operational Military Police Reports, but this idea did not prove workable. See Knapp and Campbell (2006) for a complete discussion.

[b]We explored ideas and issues associated with hands-on testing, but did not develop or administer any hands-on tests.

In an operational assessment program we expect that a multiple-choice test would comprise part of the test battery for every MOS. Prototype multiple-choice tests are already available in the form of the PerformM21 Army-wide test (Knapp & Campbell, 2006) and those developed for the Army's recent *Development of Experimental Army Enlisted Personnel Selection and Classification Tests and Job Performance Criteria* (Select21) project (Knapp,

Sager, & Tremble, 2005)[2]. Prototype multiple-choice tests were developed in this project for three MOS: 19K, 63B, and 91W. However, in some ways the work done here extends that which was done before. First, for the 19K MOS, we took the test developed as part of Select21 (Knapp et al., 2005) and added more and better graphics (e.g., animated graphics). Second, for 63B and 91W, we designed test items to measure knowledge areas, rather than tasks. For instance, for the 63B MOS, rather than having an item that asks a specific question about repairing hydraulic brakes on a particular vehicle, we developed an item that measures the test-taker's knowledge of fluid mechanics in general. Measuring knowledge areas instead of tasks makes test maintenance easier in part because it results in less frequent modifications to the test blueprint and the items when equipment is changed.

Situational judgment tests (SJTs) present job situations and pose several alternative actions that one could take to handle each situation. Respondents are asked to rate the effectiveness of each action or to select the actions they believe would be most and least effective. We developed SJTs for the 31B and 91W MOS because Soldiers in these MOS are often called on to make decisions based on situational stimuli. Research has shown that this method is effective for measuring aspects of the job that involve judgment and decision-making (McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001).

Rosenthal et al. (2005) determined that high-quality hands-on tests would likely be an integral part of an ideal assessment battery for most MOS. While relatively easy to develop, hands-on tests are quite expensive to administer and score. Therefore, although we explored some of the issues associated with operational hands-on testing for the 19K MOS in Phase II (Knapp et al., 2005), we did not develop or pilot test any hands-on tests in this research.

One method that rivals hands-on testing is high fidelity computer-based simulations. Simulations are also burdened with heavy development resource requirements, but their administration costs and requirements are typically much less than those of hands-on tests. The 14E MOS was selected in part because it seemed to be a good candidate for such a simulation. The simulation developed for this MOS was based on a scenario involving operation and maintenance of the Patriot missile system. For the 31B MOS, rather than develop a new simulation, we were able to adapt an existing simulator called the Engagement Skills Trainer (EST) 2000 to explore testing applications.

Finally, during Phase II, we also considered the implications of using available test results in lieu of new tests. This was examined in the context of the 63B MOS for which there exists relevant civilian certifications (which are used as a basis for promotion points) and the 91W MOS in which Soldiers are periodically tested and required to be certified as Emergency Medical Technicians. The Phase II report (Knapp & Campbell, 2006) discusses these implications in some detail.

[2] The objective of this project was to provide personnel tests for use in selecting and assigning entry-level Soldiers to future jobs. Development of such tests started with a future-oriented job analysis that identified the job performance requirement(s) of future first-term Soldiers and the skills, knowledges, and other personal attributes important to effective performance of the job requirements.

Test Development

Job Knowledge Tests

Most people take a multiple-choice test at some point in their lives. Each item is comprised of a stem and several (usually, three to five) response options from which to select the correct answer. Typically, there is only one correct answer to a multiple-choice test item. Some people have been exposed to other types of knowledge items including matching, ranking, or drag-and-drop (e.g., an item requiring the use of the computer mouse to "drag" labels of the parts of a drum brake system and "drop" them in the appropriate places). Computer-based testing encourages development of these non-traditional job knowledge items because they are an efficient and interesting way to present test content. As we learned in the Army-wide assessment pilot test, Soldiers felt the non-traditional items were a welcome change from typical multiple-choice items (Knapp & Campbell, 2006).

Development of a job knowledge test begins with a job analysis and proceeds through a series of steps. First, a test blueprint is prepared. A blueprint specifies (a) the total number of items to be on the test, (b) the content areas that the test will cover, (c) the number of items to be in each content area, and (d) the organization of feedback (if any) provided. Blueprints may be comprised of tasks, competencies, skills, knowledges, or any combination of these. We have noted before (Knapp & Campbell, 2006) that the Army tends to define jobs in terms of tasks, which naturally encourages the development of task-based tests. For the 19K MOS we used an existing task-based blueprint developed for the Select21 project (Knapp et al., 2005). However, we also wanted to explore developing knowledge or competency-based assessments. We felt that both the 63B and 91W MOS were ideal for this purpose.

The second step is to develop items. Item content can be developed by Army SMEs with training provided by test developers or by item developers using appropriate reference material. Either way, it is an iterative process. For this project, the items were mostly developed by project staff using various TMs, FMs, and other training material (e.g., Advanced Individual Training (AIT) training modules for the 91W MOS). We were also able to adapt items that had been developed for prior Army research projects. All items were reviewed by SMEs. Test items are often reviewed and revised several times by different SMEs to ensure they are clearly written and appropriate for testing.

Pilot testing the items is the third step. We administered the MOS-specific items to Soldiers in the Phase III pilot tests (specific sample information is presented in Chapter 3). We used the results to determine if the items (a) should be deleted from the item bank, (b) need further revision, or (c) are ready for operational use. Results from the pilot tests are discussed in Chapter 4.

An operational assessment system would require the creation of multiple equivalent test forms for each job knowledge test to allow multiple administrations and to enhance test security. The focus of this research, however, was on the development of prototype test *items*, so we did not develop multiple test forms. Instead we reviewed the item statistics and kept those items that performed well.

Prototype computer-based job knowledge tests were developed for the 19K, 63B, and 91W MOS. As previously noted, these prototype tests were not designed to completely cover the content for each MOS. Because these were computer-based tests, whenever possible, we used graphics and/or developed non-traditional items. As noted above, Soldiers liked the non-traditional items, and computers facilitate their use.

Even though the emphasis for this phase was MOS-testing, we had an opportunity to collect additional item statistics for Army-wide (common core) items developed earlier in the project. These were items that were either not piloted previously, or were piloted, but then revised. Two versions of the Army-wide test were created for Phase III: a long version (approximately 50 items) and a short version (approximately 30 items). In addition to collecting item statistics, administering a prototype common core assessment allowed us to correlate MOS and common core scores. These results are reported in Chapter 4.

Situational Judgment Tests

Situational judgment tests require examinees to evaluate alternative actions to problem scenarios (Motowidlo, Dunnette, & Carter, 1990). An SJT item presents a problem situation and several possible actions to take in each situation. The problems and actions are typically presented in text form, but may be presented via videos of actors or the use of animated characters. Examinees may evaluate the actions in several ways, such as by rating each on an effectiveness scale or by selecting the most effective and/or least effective options. SJTs are usually scored using expert judgments provided by SMEs. SJTs are realistic to the extent that problem situations (scenarios) and alternative actions (response options) are based on what actually happens or could potentially happen on the job (e.g., using critical incidents job analysis).

SJT development involves several steps. First, target performance areas must be identified (e.g., leadership, conflict management). Target performance information was collected from NCOs during the 31B and 91W proponent site visits. For the 91Ws, our review and ratings of the generalized work activities during the initial site visits suggested that interpersonal skills were important in this MOS, but not included on the task list (examples of these skills included Contributing to and Supporting Teams; Communicating with Supervisors, Peers, and Subordinates; and Establishing and Maintaining Interpersonal Relationships). Because interpersonal skills are better suited for testing by an SJT than a job knowledge test, we decided to create an SJT that would tap interpersonal skills. Similarly for the 31B MOS, SMEs at the initial site visits expressed concern that promotions were being awarded to Soldiers who were technically proficient, but lacking in interpersonal skills. Therefore, in addition to developing SJT items targeting the core 31B functions (Maneuver and Mobility Support, Police Intelligence Operations, Law and Order, Area Security, Internment, and Resettlement), we sought to develop SJT items that measured interpersonal skills specific to the military police environment.

The second step is to develop item content. SJT development requires an additional step in the job analysis phase: the generation of critical incidents. Critical incidents are actual examples of particularly effective or ineffective job performance that become the scenarios for SJT questions. During critical incident development for 91Ws, we asked SMEs to focus on job analysis-based performance requirements relating to interpersonal interactions (e.g., Contributing to and Supporting Teams, Establishing and Maintaining Interpersonal Relationships). However, the SMEs

had trouble thinking of situations that cleanly fit into these categories. Once we lifted this restriction, they were fairly prolific, developing 50 incidents during one site visit. The 31B SMEs also struggled with generating examples of interpersonal critical incidents, at least as defined by the job analysis categories. On the other hand, they had little difficulty generating critical incidents related to four of the six core 31B functions. Once the critical incidents were collected from the initial site visits, project staff edited them into SJT scenarios. At subsequent site visits, other groups of NCOs provided feedback on the scenarios and generated response options for the scenarios. Finally, HumRRO staff edited the SJT items for grammar, accuracy, realism, richness, and clarity. After final editing, there were 30 and 33 SJT items, respectively, for the 31B and 91W MOS. Item content reflects a mix of scenarios that primarily call for technical judgment and scenarios that call for such judgments in the context of challenging interpersonal contexts.

Developing a scoring key and response format comprises the third step. This required SME ratings of the effectiveness of each response option. As with the job knowledge test items, the development process for SJT items is iterative. Different SME groups were asked to provide feedback on existing scenarios and items as well as effectiveness ratings for each response option. This was done separately for each MOS (31B and 91W), and resulted in reducing the number of SJT items to 27 and 24, respectively, for the 31B and 91W MOS.

The fourth step is to pilot test the items. The draft test items were administered via computers to E4 Soldiers in the Phase III pilot test. Using the data, we finalized the test items and scoring key. Normally, this last step is to develop final test forms. However, for this research, we did not develop final test forms, but rather created a test bank of all SJT items that worked well during pilot testing. Complete results are discussed in Chapter 5. In an operational setting, it would be necessary to develop a strategy for constructing multiple equivalent forms of each SJT.

An SJT called the Leadership Exercise (LeadEx) designed for use in promotion decisions of E4 and E5 Soldiers (regardless of MOS) was developed as part of the *Maximizing 21st Century Noncommissioned Officer Performance* (NCO21) project (Knapp et al., 2002)[3]. In validation research, performance on this instrument was strongly associated with other performance measures (in particular, supervisor ratings). The LeadEx was included along with the Army-wide items during the Phase III pilot tests. This allowed us to correlate scores on both the common core and MOS-specific job knowledge tests with the LeadEx and compare these results with previous findings.

Adapting Existing Simulators and Developing New Simulations

There is considerable appeal to the idea of having dual-use technology in order to make high technology testing and training options more cost-effective. The Army has invested resources into computer-based simulators that are used for training Soldiers. For four of our target MOS, we explored the possibility of using training simulators for testing purposes. However, our research showed adapting existing simulators was not always an effective way to create test content (cf. Knapp & Campbell, 2006).

---

[3] To address the need to ensure that the U.S. Army has high quality NCOs prepared to meet the needs of the future Army, ARI initiated the project titled *Maximizing the Performance of Non-Commissioned Officers for the 21st Century (NCO21)* to examine 21st century growth decisions for NCOs. This project culminated in a set of predictor measures that were designed to improve promotion decisions for specialists/corporals (E4s) and sergeants (E5s) to the next pay grade.

For the 14E, 19K, and 91W MOS, problems included (a) too few simulators to support an operational testing program, (b) insufficient data captured in the simulators (e.g., few measures collected, focus on team rather than individual performance, limited coverage of MOS tasks), and (c) too many resources required to adapt training simulators for testing purposes. However, we were successful in developing rating scales to accompany the use of the Engagement Skills Trainer 2000 (EST 2000), a virtual weapons training system we adapted to testing 31B skills. This process is fully described in Chapter 6.

We expect our experience in PerformM21 is illustrative of what will happen with other MOS. That is, it will be difficult to identify training simulators that can be used for high stakes testing without considerable additional investment in technology enhancements (e.g., to create additional scenarios, to program the software to capture performance information, to purchase more simulators). That said, the potential for dual-use technology is very real and should be a standard consideration for each MOS testing program. A caveat, however, is that adapting simulators to serve as testing vehicles needs to be done in a manner that does not compromise their utility for training.

Developing New Simulations

As with other test types, developing a computer simulation involves a series of activities: (a) identify the critical performance areas that cannot be effectively assessed through traditional methods and gather simulation requirements; (b) develop a description of the environment to be simulated and a set of scenarios that target the identified performance areas; (c) develop story boards describing each scenario in terms of events such as user interaction, visual display changes, sounds, and user navigation; (d) design a simulation interface; (e) develop graphics, sounds, and other environmental features; (f) develop the simulation software; (g) conduct user acceptance testing and make revisions; and (h) pilot test.

We developed a fairly sophisticated simulation for the 14E MOS. We also developed two very simple simulations for the 19K MOS using animation developed for training applications. One simulation takes the 19K Soldier through the 11 steps of a .50 caliber machine gun function check and the other simulates an initiation process on the M1A2SEP tank. Along the lines of still less sophisticated simulation, we also developed a series of four 19K multiple-choice items that use animation to illustrate answers to questions regarding the correct tank formations to use under different circumstances.

Summary Comments

In Phase II we began developing a variety of MOS-specific prototype assessment items (refer to Table 2.2). With SME support we were able to approximate the processes that would be followed for an operational assessment program. While the test development process differs slightly for different test methods, it always requires a job analysis, SME input at several stages, and pilot testing. Further detail about the development of the MOS-specific prototype measures and what we learned from the process is provided in Knapp and Campbell (2006). Chapters 3 through 7 of the present report describe the Phase III pilot testing of the prototype measures. Chapter 3 describes the pilot test administration procedures and resulting samples. Chapters 4 through 6 describe results specific to the job knowledge tests, situational judgment tests, and simulations, respectively. Chapter 7 looks at relations among scores across test methods.

CHAPTER 3: PILOT TEST OF THE EXAMINATIONS

Karen O. Moriarty, Tonia S. Heffner, Jennifer L. Solberg, Kimberly S. Owens, and Charlotte H. Campbell

Introduction

This chapter concerns pilot testing of the MOS-specific prototype assessments developed in Phase II. As summarized in Chapter 2, project staff wrote new items and adapted items from previous research projects (i.e., Project A and Select21) (J. P. Campbell & Knapp, 2001; Knapp, Sager, & Tremble, 2005). SMEs from the Army and HumRRO reviewed all items. The next step, then, was to pilot test the items. This chapter provides an overview of the pilot test process and sample, with the following chapters providing results by test method.

In addition to the MOS-specific items, we administered previously developed Army-wide assessments. Army-wide assessments included the common core test developed in Phase I, and the LeadEx developed in a separate research effort (Waugh, 2004). As in Phase II, all assessments began with Soldiers providing typical demographic data on the background form and ended with the Soldiers providing feedback on areas such as:

- Computer-based testing
- Using the Digital Training Facility (DTF) for such a test
- Their test performance
- Perceived fairness (or lack thereof) of the prototype tests

Pilot Test Administration

Table 3.1 shows the types of MOS-specific tests and which version of the common core assessment was administered to the Patriot Air Defense Control Operators/Maintainers (14E), Armor Crewmen (19K), Military Police (31B), Wheeled Vehicle Mechanics (63B), and Health Care Specialists (91W). The long version of the common core test had 51 items, and the short version had 30 items. For the most part, these items were either not administered during the pilot test in Phase II, or were modified based on Phase II results and needed to be repiloted. The decision of which version of the common core assessment to administer was based on the length of the MOS tests. Table 3.1 highlights the variety of assessment types that were piloted in Phase III.

Table 3.2 shows a breakdown of the pilot test locations and the MOS tested. The "Other MOS" column refers to the long version of the common core that was administered to Soldiers who reported for testing but were not in our target MOS. These Soldiers also received the LeadEx. The results of the pilot tests are discussed individually in Chapters 4 – 6.

Table 3.1. MOS Prototype Tests

MOS	Type of MOS Test(s)	Number of Items	LeadEx SJT Administered?	Common Core Version
14E	Computer-based simulation	N/A	Yes	Long
19K	Job knowledge test Multiple choice simulation	160 3	Yes	Short
31B	Situational judgment test[a]	27	Yes	Long
63B	Job knowledge test	93	Yes	Long
91W	Job knowledge test Situational judgment test	55 24	Yes	Short

[a]The EST 2000 rating scales were administered only once to 31B Soldiers. See Chapter 6 for discussion.

Table 3.2. Phase III Pilot Tests

Date	Location	14E	19K	31B	63B	91W	Other MOS	Total
March 2005	Camp Gruber, OK			1	1	4		6
March 2005	Fort Leonard Wood, MO			23				23
April 2005	Fort Drum, NY			18	19	21		58
April 2005	Fort Hood, TX		48	10	7	10	4	79
June 2005	Fort Riley, KS		35	3	17	15	81[a]	151
June 2005	Fort Bliss, TX	70						70
July 2005	Fort Lewis, WA		1	12	5	7	95	120
August 2005	653rd Area Support Group, CA			4	6			10
August 2005	Schofield Barracks, HI			18	14	23	1	56
August 2005	336th QM Bn, OH						29	29
Aug/Sep 2005	Fort Indiantown Gap, PA		3	11	8	15		37
Sept 2005	90th Regional Readiness Command (807th MEDCOM), TX				2	28	20	50
Sept 2005	Fort Richardson, AK			11	13	13		37
Oct 2005	88th Regional Readiness Command, OH			26	1	1	14	42
Total		70	87	137	93	137	244	768

[a] Three of the Soldiers who completed the common core test also completed the 19K test.

The pilot tests were administered primarily at Army DTFs. With the exception of the 14E MOS prototype simulation, all prototype assessment items were administered via the Internet. ARI and HumRRO staff members served as test administrators (TAs). The TA function was to review the project briefing and Privacy Act statement with the Soldiers, monitor the Soldiers, and resolve any technology or computer issues that arose. Also, the TAs conducted informal interviews with Soldiers at the completion of the pilot tests to record their impressions of the test, the testing process, and general comments regarding testing.

Technology Issues

For the first part of the pilot testing, we had a repeat of some of the computer issues we had in the Phase II pilot test effort. We had computer-specific issues where one or two computers would not load graphics or would drop Soldiers from the test. We also had system-wide issues where all or nearly all of the computers in a DTF would be kicked out of the assessments or Soldiers were not able to log in. The server-provider was able to resolve most initial systemic malfunctions, and subsequently, there were very few system-wide problems.

Data Analysis and Feedback

For each prototype test (except 14E), we conducted item analyses to determine which items would be retained and included in further analyses. Once these decisions were made, we developed final scores and estimated score reliabilities. Where possible, we looked at performance differences among subgroups of examinees (i.e., subgroups based on race, gender, and deployment status). Also, we looked at MOS differences on the common core and LeadEx items. Finally, we computed correlations among the common core, LeadEx, and MOS scale scores. These results are described in the following four chapters.

Soldiers were provided with feedback on their performance on the MOS-specific, common core, and LeadEx items. As with the Phase II pilot test, the results were emailed to Soldiers who were also provided information concerning how well they did (e.g., percent correct) and how they compared to the rest of the sample (e.g., mean percent correct). Soldiers in the 14E MOS were given performance feedback on the simulation immediately after completing the simulation. We subsequently provided feedback on their performance on the common core and LeadEx items.

Overall Sample Description

Of the 768 Soldiers who participated in the pilot tests, 64 Soldiers had missing data (see Table 3.3) and had to be excluded from the analyses. The missing data were random (i.e., not related to any MOS or demographic group).We attempted to achieve equal representation among the Active, U.S. Army Reserve (USAR), and the Army National Guard (ARNG) components, but only 4% across all MOS were from the ARNG. USAR and the Active Component comprised 21% and 75% of the sample, respectively. Roughly speaking, the Active Component, ARNG, and USAR comprise 48%, 33%, and 19%, respectively, of total Army strength (Office of the Under Secretary of Defense, Personnel and Readiness, 2004). So, the sample over-represents the Active Component, under-represents the ARNG, and approximates the USAR.

Sixty-four percent of the participating Soldiers had been deployed in the previous 2 years, and of those, 70% were deployed to Iraq and 13% were deployed to Afghanistan. The overall sample was 15% female, 75% White, 15% Black, 3% Asian, and 14% Hispanic. This pattern was the same across the individual MOS: mostly White, male, and from the Active Component. Even though we requested E4 Soldiers, 20% came from other pay grades.

As shown in Table 3.4, 31B Soldiers had the most time in service (TIS) at 4.26 years, likely because this MOS had the highest percentage of Soldiers from the Reserve Components (41%). The average age of Soldiers in the overall sample was just over 24 years.

Only 13% of the sampled Soldiers had used a DTF before the pilot test. Sixty-one percent agreed or strongly agreed that DTFs were a good location for administering a test like the pilot test. Eighty percent had taken a computer-based test before, and 87% either answered "Yes" or "No Preference" when asked if they preferred computer-based tests to paper-and-pencil tests. This feedback is consistent with the feedback received in the Phase II pilot test.

Sample sizes for specific analyses reported in subsequent chapters will vary from those shown in Table 3.3. In particular, cases were dropped for some analyses where required (e.g., computing an alpha coefficient requires listwise deletion). Additionally, for each assessment, we dropped cases with too much missing data. The determination of too much missing data was made on an assessment by assessment basis, but the general rule was 30% or more.

Table 3.3. Demographic Information for Group and by MOS

Variable	14E	19K	31B	63B	91W	Other	Total
Sample Size[a]	51	94	111	76	133	239	704
Component							
Active	100%	96%	59%	73%	70%	73%	75%
USAR		1%	31%	19%	24%	26%	21%
ARNG		3%	10%	7%	6%	1%	4%
% deployed in last 2 years	31%	82%	78%	49%	63%	62%	64%
Iraq		96%	57%	68%	46%	86%	70%
Afghanistan			21%	22%	33%	2%	13%
Other	100%	4%	22%	10%	21%	12%	17%
Race/Ethnicity[b]							
Asian	2%	2%	4%	4%	5%	3%	3%
Black	16%	14%	8%	13%	17%	20%	15%
Hispanic	14%	15%	7%	14%	17%	15%	14%
White	80%	73%	84%	79%	74%	69%	75%
Other[c]	8%	6%	5%	7%	7%	11%	8%
Gender							
Male	98%	100%	80%	93%	82%	78%	85%
Female	2%	0%	20%	7%	18%	22%	15%
Pay grade							
E1 – E3	22%	17%	13%	14%	7%	11%	12%
E4	78%	67%	71%	78%	84%	88%	80%
E5 – E7		16%	16%	8%	9%	1%	8%

[a] These sample sizes differ from Table 3.2 due to missing data.
[b] Soldiers were allowed to select more than one race or ethnicity so the total percent is greater than 100.
[c] American Indian, Alaskan Native, Native Hawaiian, or Pacific Islander.

Table 3.4. Averages (in Years) on Experience-Related Variables

		Age		Time in Service		Time in Grade	
MOS	*n*	*M*	*SD*	*M*	*SD*	*M*	*SD*
14E	51	24.05	3.40	2.93	0.60	1.30	1.24
19K	94	24.73	5.43	3.69	3.16	1.39	1.22
31B	111	24.35	5.76	4.26	3.95	1.77	1.84
63B	76	24.18	5.06	3.84	3.02	1.77	1.99
91W	133	25.11	4.56	3.70	2.53	1.77	1.98
Other	239	24.42	4.73	3.47	2.23	1.72	1.50
Total	704	24.53	4.92	3.67	2.81	1.67	1.67

Summary

Seven hundred and sixty-eight Soldiers from 14 locations across both the Active and Reserve components participated in the Phase III pilot test. The pilot tests were conducted primarily at Army DTFs. We encountered technology issues similar to those in the Phase II pilot, but were eventually able to resolve them. The sample of Soldiers was primarily male, White, and from the Active component.

CHAPTER 4: JOB KNOWLEDGE TESTS

Karen O. Moriarty, Carrie N. Byrum, and Huy Le

Introduction

In this chapter we review the four prototype job knowledge tests (JKTs) administered as part of the Phase III pilot test. Chapter 2 briefly reviewed the four-step process for developing JKTs: (a) job analysis, (b) blueprint development, (c) pilot testing, and (d) creating test forms. The Phase II report documented the first two steps, and this chapter is concerned with the third step. The fourth step was not completed because it is not necessary until a test becomes operational.

Because the emphasis in Phase III was on piloting a variety of test types, only the Armor Crewman (19K) JKT comprehensively covered the applicable performance domain, using items developed as part of the Select21 project (Knapp, Sager, & Tremble, 2005). Whereas the 19K prototype test was task-based, the Wheeled Vehicle Mechanic (63B) and Health Care Specialist (91W) tests were more competency-based[4]. The fourth JKT, the common core assessment, was task-based and had two versions – a long and short version. The common core assessment was originally pilot tested in Phase II. The MOS pilot tests afforded us an opportunity to collect additional item statistics for those items which were either not administered as part of Phase II, or which were administered and subsequently edited based on the item statistics.

Chapter 3 contains the Soldier sample information for each JKT. Specifically, refer to Tables 3.3 and 3.4 for detailed demographic and background information by MOS. In the sections below we review the item selection outcomes along with descriptive statistics and reliability estimates for the scores. Also, where possible, we present the results of subgroup analyses.

Item Selection Decisions

Item selection decisions for both traditional and non-traditional items were made based on classical test theory item statistics (i.e., item difficulty, item discrimination index)[5]. For non-traditional items this requires an additional step. Standard, multiple-choice items have an item stem and four response options. One of those options is the correct response (the key), and the rest are distracters. Figure 4.1 is an example of a matching item. The stem asks Soldiers to match each stimulus (e.g., nasal cavities) with a response option shown in the drop-down box.

Item statistics for non-traditional items allow us to look at the overall item performance, or at each stimulus (see Figure 4.1). That is, we can calculate an item discrimination and item difficulty index for the overall item, and for each stimulus, we can look at item-total correlations and response distributions. We learned in Phase II (Knapp & Campbell, 2006) that non-traditional items have better item-level statistics than traditional items. If one thinks of each stimulus and drop-down box as a separate item, the reason is clear. These non-traditional items have more "data points." Just as

[4] Throughout this report, we use the term "competency-based" to refer to test content intended to capture the knowledge base underlying successful task performance.

[5] The scoring procedure for non-traditional items differs from that for traditional items. This is briefly reviewed below. For a more comprehensive discussion, refer to the Phase II technical report (Knapp & Campbell, 2006).

increasing the number of items on an assessment will usually increase reliability, increasing the number of stimuli in an item will usually increase the item-total correlation.

Each assessment developer reviewed these statistics and decided which items to keep and which to drop. The goal was to maximize test reliability and blueprint coverage. Below we briefly review the outcomes of these decisions for each JKT.

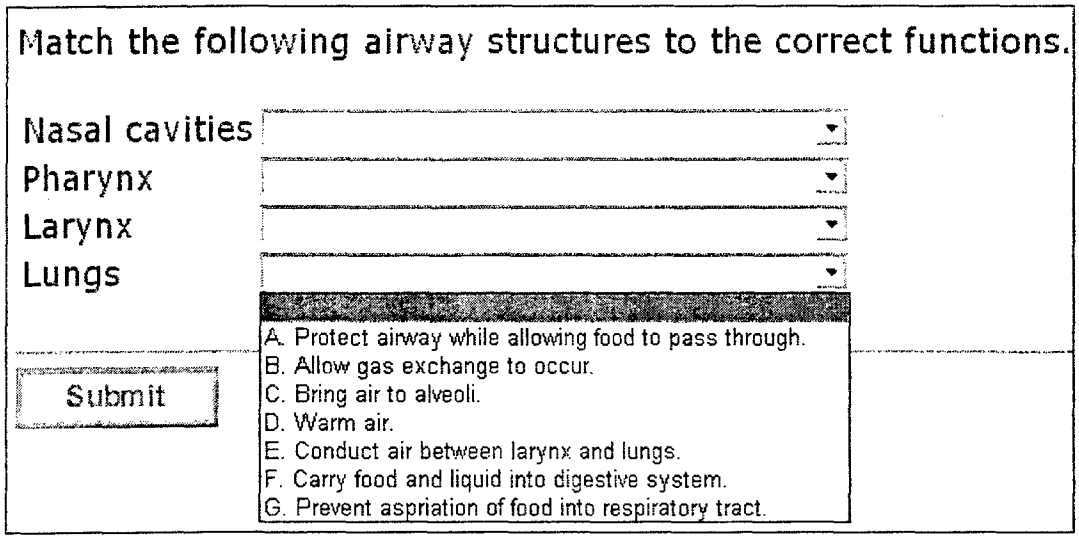

Figure 4.1. Screen shot of a matching item.

19K MOS Job Knowledge Test

For the 19K JKT, an item blueprint that had already been developed as part of Select21 (Knapp et al., 2005) was used as the basis for the 19K prototype test administered during this phase. Table 4.1 shows the distribution of items that were pilot-tested and the distribution of those items that were retained. There were four categories with low retention rates: Evacuate Wounded, Load and Unload Tank Main Gun, Tank-Mounted Mine Clearing, and Tank Recovery Functions. Inspection of the response distributions for the dropped items in the four categories suggested that the Soldiers were guessing when choosing a correct response. Closer investigation of the content of the items suggested reasonable explanations for this finding.

First, of the three items dropped in the Evacuate Wounded category, two dealt with the tasks involved in positioning Soldiers during an evacuation from a driver's hatch. It is possible that Soldiers had not received sufficient task training or these tasks were not trained, trained improperly, or not trained recently. Also, it is important to note that this category was initially comprised of only five items. As such, the low retention rate may be item content sampling error.

Second, for the Tank-Mounted Mine Clearing category, the Soldiers may have been hindered by lack of availability of equipment for training on these tasks. Access to the mine clearing apparatus varies by unit and by location resulting in many Soldiers accumulating little applied experience with this equipment. In fact, during item development, SMEs commented that the tasks covered by this category are ones on which many units train their Soldiers on an as-needed basis due to the scarcity of equipment and the infrequency with which the tasks are required.

Table 4.1. 19K Prototype Item Distribution

Area	Category	Original	Final	% Kept
Tank Gun Ammunition	Inspect Ammunition	7	5	71%
	Stow Ammunition	6	6	100%
Tank Machine Guns	Tank Machine Guns	22	18	82%
SINCGARS	Operate SINCGARS in Net-Centric Environment	9	7	78%
Tank Crew Functions	Evacuate Wounded	5	2	40%
	Extinguish Fire	4	4	100%
	Use Visual Signaling Techniques	6	4	67%
Tank Driver Functions	Drive Tank	24	19	79%
	Perform BDAO* Checks	7	5	71%
	Prepare Driver's Station	10	6	60%
	Start & Stop Tank	8	5	63%
Tank Loader Functions	Load & Unload Tank Main Gun	7	2	29%
	Main Gun Functions	8	4	50%
Tank Maintenance Functions	Main Gun Maintenance	7	5	71%
	Prepare Powerpack Removal	4	3	75%
	Remove/Install Track Block	7	5	71%
	Replace Thrown Track	3	3	100%
Tank-mounted Mine Clearing	Tank-Mounted Mine Clearing	6	0	0%
Tank Recovery Functions	Tank Recovery Functions	13	5	38%
	TOTAL	163	108	66%

*BDAO = before, during, and after operations

Third, items included in the Tank Recovery Functions category appeared, in hindsight, to have captured more unusual recovering operations such as multiple tank recovery and overturned tank. These types of operations are not performed regularly in units. As such, Soldiers' ability to recall the doctrine related to tank recovery may have been hampered by their on-the-job experience. Given that Soldiers were given no time to prepare for the test, it is reasonable to believe that they may have forgotten the doctrine related to this category.

It is less clear why Soldiers were inclined to guess on the Load and Unload Tank Main Gun category. The tasks covered by this category are ones on which Soldiers receive regular training. Soldiers are not hindered by lack of access to the necessary equipment. Moreover, all 19K Soldiers, regardless of rank, should be familiar with, and have applied experience in, the Loader position. However, because the majority of the sample came from a single installation, it is possible that item responses in the Load and Unload Tank Main Gun category reflect characteristics of this particular group of 19K Soldiers. Moreover, a review of the content of the items dropped from this category revealed that two of the items dealt with night vision viewer equipment, while the remaining three dropped items covered more general Loader tasks. As such, this sample of Soldiers may have had different experiences with the night vision viewer and, similar to the Tank Recovery Functions category, may have experienced training on the general Loader tasks that is unlike the training found in doctrine.

63B MOS Job Knowledge Items

To guide our item development efforts, we conducted a competency-based blueprint exercise based on the topics covered in the TM 9-8000 *Principles of Automotive Vehicles*. The SMEs advised us that this manual is the basis for Advanced Individual Training (AIT). TM 9-8000 has three levels of topics, ranging from general to specific. A sampling of this layout is shown in Table 4.2. We created a blueprint for these three categories using a series of SME-generated weights and ratings. Our blueprint analyses suggested we focus our limited resources on Electrical Systems and Engines, which accounted for 81% of the points. We also developed a few items concerning brakes.

Table 4.2. 63B Blueprint Categories Sample

Category 1	Category 2	Category 3
Engines	Gasoline Fuel Systems	Principles of Carburetion Fuel Injection Systems
	Diesel Fuel Systems	Combustion Chamber Design Timing Device
Electrical Systems	Charging Systems	AC Generator Systems DC Generator Systems
	Basic Principles of Electricity	Electrical Measurements Semi-Conductor Devices

The 63B MOS began with 93 items, as shown in Table 4.3. A review of the item statistics along with item content revealed that some items were simply too theoretical, which was a concern of project staff from the start[6]. These items were dropped, which accounts for some of the low retention rates. The reference materials, which are used in AIT, included TM 9-8000 and various technical manuals for common vehicles such as the HMMWV. These reference materials review and discuss theoretical concepts, which SMEs told us were "fair game."

Table 4.3. 63B Prototype Item Distribution

Area	Category	Original	Final	% Kept
Brakes	Brakes	8	4	50%
Electricity	Basic Principles	12	9	75%
	Charging Systems	9	5	55%
	Repair Wiring	4	1	25%
	Starting Systems	14	5	36%
Engines	Conventional Engines	14	11	79%
	Diesel Fuel Engines	14	6	43%
	Engine Cooling Systems	16	8	50%
	Miscellaneous	2	2	100%
	TOTAL	93	51	55%

[6] During test sessions, many 63B Soldiers expressed displeasure that the electrical systems items were on the test. They commented that they were not very good with this topic.

We developed more declarative knowledge than applied knowledge items. Items measuring declarative knowledge have a place in JKTs. However, we would prefer a greater number of knowledge application questions because these items more closely simulate real working conditions. Unfortunately, that requires much more SME involvement than we were able to secure in this project.

For this project, project staff developed items using materials provided by the SMEs. These items were then subject to SME review. In an operational program, item relevancy is not likely to be a such a significant problem because it is envisioned that SMEs, such as retired NCOs, would develop items under the tutelage of assessment development experts. SMEs in such technically-oriented MOS are better able to create realistic knowledge application items than are technically-unknowledgeable item development experts.

91W MOS Job Knowledge Items

As with the 63B JKT, we developed a competency-based blueprint for the 91W JKT. However, the method used differed from that described above. We first conducted a task-based blueprint exercise using the 91W task list. Then, with SME help and guidance, we determined those competencies that underlay the tasks that were rated most highly. This process led us to emphasize those areas shown in Table 4.4. Items that were specifically designed to cover competencies rather than tasks are in the Airway and Circulation areas. Note the higher retention rate among these items. Approximately 50% of the items are non-traditional, which certainly helped the retention rate, because, as noted earlier, non-traditional items have better item-level statistics than traditional items. We did not experience the same difficulty with these competency-based items as with the 63B items. We believe this may be due to the fact that we were limited, because of the reference material provided, in how theoretical we could get with the 91W items. The reference material, which is used in AIT, focused mostly on anatomy and function.

Table 4.4. 91W Prototype Item Distribution

Area	Category	Original	Final	% Kept
Airway	Airway	10	9	90%
Circulation	Circulation	10	10	100%
Sterile Dressings	Sterile Dressings	9	4	44%
Manage IVs	Initiate IVs	7	1	14%
	Manage Patient with IVs	4	4	100%
Measure & Record Vital Signs	Measure Blood Pressure	3	1	33%
	Measure Pulse	6	3	50%
	Measure Respiration	2	2	100%
Triage & Evacuation	Triage & Evacuation	4	4	100%
	TOTAL	55	38	69%

The Sterile Dressings, Manage IVs, Measure and Record Vital Signs, and Triage and Evacuation areas consisted of updated Project A items (J. P. Campbell & Knapp, 2001). These areas, for the most part, did not have as high item retention rates as Airway and Circulation.

There was quite a bit of disagreement among SMEs during item development and review about which options were correct for items from the Initiate IVs category. We attempted to resolve this by revising the response options to reduce ambiguity. However, the 14% retention rate in this area suggests we were not successful.

In the Triage and Evacuation area, we created a multiple-choice "simulation" where each Soldier was presented with the same first three items. The first item required Soldiers to sort five patients with various injuries into triage categories. The second item required sorting the same patients into evacuation categories. The third item then stated that while evacuation to a forward support unit had been called, the vehicle that arrived only has room for one patient. Soldiers were required to select which of the five patients should go. The fourth item, which was similar to the third item in that it also required Soldiers to decide which patient to evacuate, was determined by the answer to the third item. As can be seen, this area performed well. It should be noted that the first two items in this area are non-traditional, matching items.

Common Core Job Knowledge Tests

All of the items on the common core short form are on the long form. Although the name implies complete prototype tests, neither form was intended to represent the entire common core domain because our goal was to collect item statistics on unused or revised items rather than to conform to the blueprint. We piloted what we believed to be our best common core items during Phase II. In Phase III, the remaining items that had been developed as well as those revised from Phase II were piloted. For this reason, it is not surprising that the percent of items retained was low (see Table 4.5).

Table 4.5. Common Core Item Distribution

		Original		Final		% Kept	
Area	Category	Long	Short	Long	Short	Long	Short
Skill Level 1	Combat Techniques	6	6	3	3	50%	50%
	First Aid	5	4	3	2	60%	50%
	Navigate	5	0	3	0	60%	
	NBC	2	2	2	2	100%	100%
	Weapons	6	2	2	2	33%	100%
Skill Level 2	Combat Techniques	2	2	1	1	50%	50%
	First Aid	3	2	0	0	0%	0%
History/Values	Courtesy & Customs	5	3	2	1	40%	33%
	Values	1	1	1	1	100%	100%
	Volunteer Army	1	0	1	0	100%	
Leadership	Chain of Command	2	1	1	1	50%	100%
	Troop Leading Procedures	2	1	1	0	50%	0%
	Risk Management	2	1	1	1	50%	100%
	Principles of Discipline	2	1	2	1	100%	100%
Training	Roles & Responsibilities of NCO	2	0	1	0	50%	
	Train Subordinates	4	4	1	1	25%	25%
	Preparatory Marksmanship Training	1	0	1	0	100%	
	TOTAL	51	30	26	16	51%	53%

Descriptive Statistics and Reliability Estimates

The items that were retained from each JKT were used to calculate statistics for each test and sub-domain. Recall that non-traditional items were liberally used in creating the JKTs. These items are usually worth more than 1 point, which is why columns for both the number of items and number of points (columns three and four, respectively) are included in Tables 4.6, 4.7, and 4.8. Weights for the non-traditional items were derived using a procedure explained in the Phase II report (Knapp & Campbell, 2006). The principle behind this procedure is that there are three ways to think about the "worth" of a non-traditional item. Let us assume there is a drag and drop item with five pieces to be dragged and dropped into their correct locations. First, the item could be worth 5 raw points – one for each piece that is correctly dragged and dropped. This may overweight the item relative to the traditional multiple-choice items. Second, the item could be worth 1 point – credit given only for correctly dragging and dropping all five pieces. This may under-value the item relative to a traditional item. Third, one could empirically determine a maximum weight for the item that reflects its informational value. We used the third method. So, it is important to note that the scores reported below represent percentage of *points* correct – not items.

Table 4.6. Descriptive Statistics for the 19K JKT

					Percent of Points Correct			
Scale	*n*	Number of Items	Number of Points	Reliability	Min	Max	*M*	*SD*
Tank Gun Ammunition	70	11	19	0.807	0.00	90.79	61.25	21.32
Tank Machine guns	63	18	19	0.692	12.63	85.79	54.50	16.55
SINCGARS	71	7	7	0.574	0.00	100.00	44.14	25.38
Tank Crew Functions	69	10	12	0.684	0.00	100.00	51.97	24.22
Tank Driver Functions	67	35	35	0.817	17.14	91.43	65.31	17.06
Tank Loader Functions	77	6	6	0.570	0.00	100.00	67.00	25.04
Tank Maintenance Functions	70	16	18	0.815	0.00	100.00	57.12	25.43
Tank Recovery Functions	77	5	5	0.184	0.00	100.00	49.39	22.58
Total	38	108	121	0.935	21.01	87.75	58.92	15.82

Table 4.7. Descriptive Statistics for the 63B JKT

					Percent of Points Correct			
Scale	*n*	Number of Items	Number of Points	Reliability	Min	Max	*M*	*SD*
Brakes	58	4	12	0.508	0.00	100.00	77.35	24.44
Electrical System	42	20	24	0.459	21.21	74.17	49.94	11.71
Engines	64	25	37	0.836	23.15	98.20	68.15	16.88
Miscellaneous	71	2	2	0.454	0.00	100.00	60.56	39.57
Total	36	51	75	0.865	21.53	88.78	63.71	14.03

We computed coefficient alpha reliability estimates for both the total scores and subscores. Because most of the subscores are based on relatively small numbers of items, these results should be interpreted with caution. The Electrical System subscore includes 20 items, but still has low reliability (see Table 4.7) suggesting the 20 items measure multiple constructs. With the exception of the common core assessments, the total scale reliability estimates are high (i.e., .80 or higher). Overall we are pleased with these numbers, considering that, except for the 19K JKT, these assessments have less than the ideal number of items.

Table 4.8. Descriptive Statistics for the 91W JKT

Scale	*n*	Number of Items	Number of Points	Reliability	Percent of Points Correct			
					Min	Max	*M*	*SD*
Airway	129	9	12	0.630	4.2	79.2	50.3	16.4
Circulation	130	10	16	0.630	13.5	87.5	57.0	16.4
Manage IVs	136	5	5	0.332	20.0	100.0	84.1	18.6
Vital Signs	127	6	7	0.500	21.4	100.0	80.8	16.7
Sterile Dressings	128	4	4	0.438	0.00	100.0	73.4	25.8
Triage & Evacuation	121	4	11	0.518	31.4	94.3	70.9	13.5
Total	105	38	55	0.804	27.5	86.1	66.6	9.6

The second column in Tables 4.6 through 4.8 indicates the sample size used to generate the reported statistics. Reliability estimates require listwise-deletion for missing data for each calculation. Because of this, the sample sizes vary from scale to scale. However, for purposes of providing Soldier feedback, we estimated scores for all Soldiers who did not have more than 30% missing data.

MOS Job Knowledge Tests

Table 4.6 shows that the 19K JKT performed quite well. The estimated reliability is .94, and the total score ranges from 21% to 88% correct. The 19K MOS is closed to women, so gender analyses could not be performed. We adopted 20 as our minimum sample size for subgroups. Since there were not enough minorities to satisfy this requirement, no race or ethnic subgroup comparisons could be completed.

Overall, the 63B prototype assessment performed well with a reliability estimate of .87. The range for the total score was approximately 22% to 89% correct. As with the 19K MOS, we did not have enough minorities or females to conduct subgroup analyses.

The range for the total score for the 91W job knowledge items was similar to the 19K and 63B tests – 28% to 86%. The estimated reliability was a little lower than the other MOS tests, at .80. This MOS provided our largest sample so we were able to perform subgroup analyses (see Table 4.9). The race/ethnicity results are what one would expect, based on the literature in high-stakes testing in employment and education (Sackett, Schmitt, Ellingson, & Kabin, 2001). That is, White Soldiers scored higher than Black and Hispanic Soldiers. Although this would lead us to expect the JKTs to exhibit race differences in an operational setting, the effects may well be reduced when Soldiers have the opportunity to prepare for the tests.

Male Soldiers performed better than female Soldiers on the 91W test. If this performance difference is not simply attributable to sampling error or differences in sample sizes (which it may well be), this finding is difficult to explain. Males and females are equally likely to be in TDA or TOE units, and 68% are White, so neither unit assignment nor race provides an explanation. Historically, males have outperformed females on standardized tests, particularly in mathematics and science (Bridge, Judd, & Moock, 1979; Jencks, 1972). However, efforts in the 1990s by parents and teachers have reduced this gap, and in some cases, reversed it (Whitmire, 2006).

During item development, the SMEs suggested that the 91W JKT items were more appropriate for Soldiers in a hospital setting (TDA unit), and therefore felt Soldiers in a field setting

(TOE unit) would be disadvantaged. On the 91W background information form we asked Soldiers to indicate to which unit type they were assigned. In contrast to what the SMEs anticipated, the TOE Soldiers actually performed slightly better than the TDA Soldiers, although the difference was not significant.

Table 4.9. Subgroup Differences in the 91W JKT Scores

	n	*M*	*SD*	Effect Size	*p*
Unit Type					
TDA (Hospital)	26	64.90	15.79	-0.17	0.08
TOE (Field)	110	67.36	14.77		
Gender					
Female	26	55.16	12.40	-0.73	0.001
Male	110	63.20	11.03		
Race[a]					
Black	25	55.61	13.17	-.71	.003
White	96	63.34	10.89		
Hispanic	24	59.36	10.65	-.37	0.11
White	83	63.45	11.02		
Black	25	55.61	13.17	-.35	0.25
Hispanic	24	59.36	10.65		

Note. Effect sizes are calculated as (Mean of non-referent group – Mean of referent group)/*SD* referent group. Referent groups are the second category listed within each pair (e.g., Male, White).

[a] Soldiers were allowed to select more than one race/ethnicity, which is reflected in the varying group sizes.

Common Core Job Knowledge Tests

For the common core JKTs, we did not create subscale scores as was done for the other JKTs. The estimated total score reliabilities are low, but not unexpected because of the few number of items, and the diverse topics, comprising each form. Neither the long nor the short version was intended to represent a complete assessment. Each form simply provided a means for us to collect additional item statistics. In creating each MOS test battery, we estimated that Soldier assessment would take longer than it did (see Table 3.1). Had we known that administering the MOS assessments would have required less time, we would have made the common core assessments longer and, thus, more representative of a complete assessment.

Table 4.10 shows that the common core sample sizes are larger than the MOS JKTs. This is because, as shown in Table 3.1, each MOS pilot test battery included either the short or long form of the common core assessment. Additionally, Soldiers who were tasked to participate in the pilot test, but were not assigned to any of our target MOS, completed the long form of the common core assessment.

Because a short form score of the common core JKT could be computed for all examinees (including those administered the long form), we compared subgroups using the short form score. The subgroup analyses effect sizes are smaller than those from the 91W MOS JKT.

These results are probably also more generalizable to the population of examinees because the subgroup sample sizes were large enough to yield more stable estimated effect sizes.

Table 4.10. Descriptive Statistics for the Common Core Items

					Percent Correct			
	n	Number of Items	Number of Points	Reliability	Min	Max	*M*	*SD*
Long Version	372	26	36	0.648	22.00	100.00	63.66	13.71
Short Version	540	16	22	0.556	14.00	100.00	66.41	15.15

Table 4.11 Subgroup Differences in the Common Core (Short form) JKT

	n	*M*	*SD*	Effect Size	*p*
Deployed Recently					
No	173	64.90	15.79	-0.17	0.08
Yes	347	67.36	14.77		
Gender					
Female	71	60.49	14.35	-0.47	0.00
Male	430	67.49	15.04		
Race[a]					
Black	69	61.65	13.92	-.40	0.00
White	388	67.70	15.11		
Hispanic	49	63.65	14.50	-.28	0.07
White	462	67.92	15.30		
Black	69	61.65	13.92	-.32	0.07
Hispanic	60	66.13	13.98		

Note. Effect sizes are calculated as (Mean of non-referent group – Mean of referent group)/*SD* referent group. Referent groups are the second category listed within each pair (e.g., Male, White).

[a] Soldiers were allowed to select more than one race/ethnicity. In instances where a Soldier was a member of both groups (i.e., White and Black), he was assigned to the minority group for analyses.

We were also able to compare MOS performance on the short form of the common core test. Table 4.12 shows the descriptive statistics. There were two significant differences. Both the 91W and 31B Soldiers scored significantly higher than the 19K Soldiers. The effect sizes (*d*) were .69 and .46, respectively. The effect sizes were calculated as the differences between the means of the two groups divided by the pooled standard deviation. The 14E Soldiers were not included in the subgroup comparisons because of their small sample size.

Table 4.12. Common Core Performance by MOS

MOS	*n*	*M*	*SD*
91W	107	70.45	13.54
14E	19	69.26	18.46
31B	93	67.30	13.52
63B	57	64.49	11.46
19K	51	59.94	20.45

Correlations Between Common Core and MOS-Specific JKT Scores

The highest correlation was between the common core and the 19K MOS scores at r = .72 (n = 45), p = .001. This might be influenced by three factors. First, the 19K MOS is a combat arms (CA) MOS, while 63B and 91W are combat service support (CSS) MOS. CA MOS are more likely to have more of their tasks and knowledges overlap with common core tasks and knowledges. Second, the 19K JKT was more reliable than the 63B or 91W JKTs (.94 compared to .87 and .80, respectively). Third, the 63B and 91W JKTs represent only a narrow portion of the performance domain of their respective MOS. The second highest correlation was that between the common core and the 63B scores, r = .40 (n = 57), p = .002, and the smallest correlation was that between the common core and 91W scores, r = .39 (n = 108), p = .0001. We believe these three correlations are attenuated to some extent due to the low reliability of the common core short form.

Soldier Reactions to JKTs

Soldiers were asked to provide feedback on the tests and testing process verbally in informal interviews and in a survey as part of the Internet-based pilot test. In the informal interviews, Soldiers gave mostly positive feedback on both the tests themselves and the testing process. For the most part they felt the tests were fair and needed. They also liked the liberal use of non-traditional items. In the online feedback surveys Soldiers were asked for their impressions using the following types of questions:

- "Effective" questions were phrased, "Imagine you had all the time you needed to prepare for this test. How effectively do you think the test would measure your knowledge of _____?"
- "Well" questions were phrased, "How well do you think you did on the _____ items?"

They were asked these two questions for each subscale (e.g., Brakes, Electrical Systems, and Engines for the 63B MOS) using a 5-point rating scale ranging from 1 for Very Poorly to 5 for Very Well.

The patterns of responses across all of the JKTs are very similar to the data obtained in the Phase II pilot test (Knapp & Campbell, 2006). Soldiers' responses to the "effective" questions indicated that most felt the tests would do well or very well in measuring their knowledge even though their responses to the "well" questions indicated they did not feel they scored well on the tests.

Of the five MOS that we researched in the project, the 19K MOS SMEs were the most opposed to JKT testing. They were very concerned about having Soldiers with "book smarts" but without "common sense" or "street smarts." This concern is confirmed by the pattern of responses to the "effective" and "well" questions (see Tables 4.13 and 4.14). Responses to the "effective" questions are more negative for 19K than the other JKTs, and the gap in well and very well responses between the "effective" and "well" questions is smaller.

In terms of how well 19K Soldiers think they did, Tank Driver Functions is clearly the area on which they felt they did best. They indicated they felt they did the worst on SINCGARS, Tank Recovery Functions, and Tank Maintenance Functions.

Table 4.13. 19K Effective Questions Responses

Scale	Very Well	Well	Neither Well nor Poorly	Poorly	Very Poorly
Tank Gun Ammunition	19%	38%	27%	12%	4%
Tank Machine guns	11%	44%	26%	11%	8%
SINCGARS	8%	41%	28%	22%	1%
Tank Crew Functions	15%	43%	26%	12%	4%
Tank Driver Functions	16%	41%	27%	13%	3%
Tank Loader Functions	13%	41%	31%	11%	4%
Tank Maintenance Functions	12%	42%	28%	10%	8%
Tank Recovery Functions	9%	47%	30%	9%	5%

Note. $n = 74$.

Table 4.14. 19K Well Questions Responses

Scale	Very Well	Well	Neither Well nor Poorly	Poorly	Very Poorly
Tank Gun Ammunition	12%	35%	34%	16%	3%
Tank Machine guns	7%	36%	36%	18%	3%
SINCGARS	2%	24%	46%	24%	4%
Tank Crew Functions	8%	38%	35%	15%	4%
Tank Driver Functions	10%	42%	31%	16%	1%
Tank Loader Functions	7%	34%	38%	20%	1%
Tank Maintenance Functions	5%	24%	45%	19%	7%
Tank Recovery Functions	3%	23%	38%	28%	8%

Note. $n = 74$.

Tables 4.15 and 4.16 contain the response data for the 63B MOS. Of note is the relatively poor standing of the Electrical System scales. However, given the previous discussion about the highly theoretical nature of many of the Electrical System items, this result is not too surprising.

One of the ATPAT members requested that we ask 63B Soldiers where they acquired the knowledge to answer the test questions. Eleven percent indicated that the knowledge came mostly or nearly all from the schoolhouse, whereas 62% indicated it came mostly or nearly all from the field.

Table 4.15. 63B Effective Questions Responses

Scale	Very Well	Well	Neither Well nor Poorly	Poorly	Very Poorly
Brakes	27%	53%	16%	3%	1%
Electrical System	27%	40%	26%	4%	3%
Engines	20%	59%	14%	6%	1%

Note. $n = 68$.

Table 4.16. 63B Well Questions Responses

Scale	Very Well	Well	Neither Well nor Poorly	Poorly	Very Poorly
Brakes	13%	38%	36%	13%	0%
Electrical System	4%	6%	43%	35%	12%
Engines	9%	43%	38%	9%	1%

Note. $n = 68$.

With the 63B JKT we experimented with providing embedded links to electronic troubleshooting charts for many of the items. Because of technological issues we could not include the entire charts, many of which were 20 or more pages. Instead, we included information we felt was most relevant to answering the question. Sixty-seven percent of the Soldiers indicated that they attempted to access the troubleshooting charts, and, of those, 88% indicated they were somewhat or very helpful. Discussions with the Soldiers indicated that most of the problems were because the charts did not contain all of the information the Soldiers were expecting to see, which, given adequate resources and bandwidth, is an easy fix. Only 16% of the Soldiers indicated a preference of paper to electronic manuals. This is encouraging given this MOS's move to more electronic and fewer paper manuals.

The 91W SMEs expressed some concern about competency testing (see Tables 4.17 and 4.18). Some of this concern was related to "book" versus "street" smarts, but primarily they believe they are adequately tested between the requirements to maintain a current EMT license and the Semi-Annual Combat Medic Skills Verification Test (SCAMS-VT) (see Knapp & Campbell, 2006 for complete discussion). There is also the issue that this MOS has a very strong haptic skill requirement. It is one thing to recognize that a certain injury requires the insertion of a chest tube, but it is quite another to efficiently and correctly insert that tube. So, the favorable responses to the "effective" questions are a positive sign.

Table 4.17. 91W Effective Questions Responses

Scale	Very Well	Well	Neither Well nor Poorly	Poorly	Very Poorly
Airway	28%	48%	16%	6%	2%
Circulation	27%	48%	18%	6%	1%
Manage IVs	24%	51%	18%	6%	1%
Vital Signs	24%	43%	22%	10%	1%
Sterile Dressings	26%	52%	18%	3%	1%
Triage & Evacuation	25%	50%	16%	8%	1%

Note. $n = 135$.

Some Soldiers, in discussing their impressions of the common core assessment, mentioned that the items did not seem to "go together" well. Indeed, as noted previously, these scales are the most incomplete of all the JKTs. While the results in Tables 4.19 and 4.20 are not as positive as the Phase II results, they are encouraging.

Table 4.18. 91W Well Questions Responses

Scale	Very Well	Well	Neither Well nor Poorly	Poorly	Very Poorly
Airway	10%	46%	32%	11%	1%
Circulation	10%	41%	39%	7%	3%
Manage IVs	18%	57%	21%	3%	1%
Vital Signs	15%	47%	32%	4%	2%
Sterile Dressings	9%	45%	33%	10%	3%
Triage & Evacuation	13%	49%	29%	9%	0%

Note. $n = 135$.

Table 4.19. Common Core Effective Questions Responses

Scale	Very Well	Well	Neither Well nor Poorly	Poorly	Very Poorly
Common Tasks	18%	45%	28%	6%	3%
Army/NCO History	17%	44%	28%	9%	2%
Leadership	18%	46%	27%	7%	2%
Training	16%	46%	30%	6%	2%
Army Values	31%	41%	22%	4%	2%

Note. $n = 664$.

Table 4.20. Common Core Well Questions Responses

Scale	Very Well	Well	Neither Well nor Poorly	Poorly	Very Poorly
Common Tasks	7%	37%	43%	10%	3%
Army/NCO History	6%	22%	45%	23%	4%
Leadership	6%	34%	46%	11%	3%
Training	7%	34%	48%	9%	2%
Army Values	22%	42%	31%	3%	2%

Note. $n = 664$.

Discussion and Recommendations

JKTs are a relatively easy and efficient way to measure Soldier proficiency. First, the SME investment is not as high as for simulations or SJTs. Second, they can be developed to cover most competencies or tasks, although the measurement of physical skills or specialized judgment is better suited to other types of tests or assessments. Third, they fit well into any MOS assessment strategy.

We have shown the benefits of adding more non-traditional items and graphics to the standard, multiple-choice test. One benefit noted is positive Soldier reaction. Many Soldiers said they welcomed the break from traditional multiple-choice items. A second benefit is efficiency in content presentation. Digital and/or color graphics reduce the reading requirement, and non-traditional items allow multiple knowledge points to be measured with one item.

Although non-traditional items reduce the reading requirements for these tests, test scores still showed evidence of subgroup differences. The differences observed here for the MOS tests are based on small minority group sample sizes, so the findings should be interpreted with caution. The findings are, however, consistent with research with high stakes testing. That is, even with various interventions (e.g., coaching, low reading level test items), Black and Hispanic examinees are likely to have lower average scores than White examinees (Sackett et al., 2001). Well-constructed tests with high "face validity," such as those developed here, have been shown to at least somewhat reduce subgroup differences, but some differences can still be expected.

There has been a lot of discussion in ATPAT meetings about whether to include MOS-specific and/or Army-wide tests as part of a competency system. The major concerns are resources: both financial and time. The correlations reported in this chapter suggest that although the MOS and common core assessments are significantly correlated, they are clearly capturing different portions of the Soldier performance domain. This supports including both types of tests if resource issues can be adequately addressed.

CHAPTER 5: SITUATIONAL JUDGMENT TESTS

Jennifer L. Burnfield, Gordon W. Waugh,
Andrea Sinclair, Chad Van Iddekinge, and Karen O. Moriarty

Introduction

This chapter reviews the tests results of three situational judgment tests (SJTs) administered in the Phase III pilot tests. The Army-wide Leadership Exercise (LeadEx) was developed in an earlier Army project (Waugh, 2004). We developed pilot versions of two MOS-specific SJTs, one for Military Police (31B) and the other for Health Care Specialists (91W). Although similar in nature, the LeadEx and the MOS-specific SJTs use different response formats. On the LeadEx, examinees identify the response option (out of four choices) they think would be most effective and the response option they think would be least effective for addressing the problem. On the MOS SJTs, examinees rate the effectiveness of each response option on a 7-point scale, and are allowed to assign the same rating to multiple response options. Sample items using each of these response formats are shown in Figure 5.1. On both types of SJTs, the scoring key was developed using the effectiveness ratings of SMEs. Specifically, the *keyed effectiveness* for each response option is the mean SME rating.

Sample "Most and Least" Choice Format

Instructions: For each item, mark which course of action you would be MOST likely to follow with an "M" and mark the choice you would be LEAST likely to choose with and "L"

As a junior NCO, you need to counsel a subordinate. What would be your priority when preparing for and conducting the counseling?

___ a. Prepare a course of action that you want the Soldier to follow
___ b. Plan to guide and encourage the Soldier to arrive at his own solutions
___ c. Focus on the sanctions and rewards that you control
___ d. Follow the outline of the DA for 4856-R, General Counseling Form

Sample Effectiveness Rating Format

One of your fellow Soldiers feels like he does not have to pitch in and do the work that you were all told to do. What should you do?

Rate the effectiveness of each response option based on the scale below.

	1	2	3	4	5	6	7
Explain to the Soldier that he is part of a team and needs to pull his weight.	O	O	O	O	O	O	O
Report him to the NCO in charge.	O	O	O	O	O	O	O
Keep out of it; this something for the NCO in charge to notice and correct.	O	O	O	O	O	O	O
Find out why the Soldier does not feel the need to pitch in.	O	O	O	O	O	O	O

Ineffective Action — The action is likely to lead to a bad outcome

Moderately Effective Action — The action is like to lead to a passable or mixed outcome

Very Effective Action — The action is likely to lead to a good outcome

1 2 3 4 5 6 7

Figure 5.1. Sample situational judgment test items.

Development of LeadEx Scores

As mentioned, the 24-item LeadEx was developed in a prior research effort (Waugh, 2004). Therefore, the items were scored using the key and scoring algorithm developed in that research. Specifically, the score for each item was the keyed effectiveness of the option picked by the respondent as *most* effective minus the keyed effectiveness of the option picked by the respondent as *least* effective. It is important to note that, because the LeadEx uses a different response format and scoring approach from the 31B and 91W SJTs, the scores are not on comparable metrics. The scores for the LeadEx represent the percentage of possible points earned.

Development of MOS SJT Scores

The MOS-specific SJT scores were developed in a two-step process as a part of the Army's Select21 project (Waugh & Russell, 2005). The first step entailed reviewing pilot scenarios (and options within scenarios) to determine which should be retained. Here, we adopted the standard of retaining four response options for each item. In step two, scores were derived by comparing the Soldier's effectiveness rating for each option to the mean rating obtained from expert judges (SMEs).

Selection of Items and Response Options

The pilot test form of the 31B SJT had 27 items, and the 91W SJT had 24 items. Each item had four to seven response options. Both rational and empirical methods were used to select the final set of items and response options. In terms of rational methods, item content was examined for redundancy in the scenarios and options. With respect to empirical methods, the following rules were used to decide which options and items to drop:

- The highest and lowest keyed effectiveness values among an item's options must be at least 2.0 (approximately).
- The standard deviation among the SMEs' ratings for an option must be less than 2.00.
- If more than four options within an item survived the first two rules of thumb, then we retained the set of four options that were spread out the most (in terms of their keyed effectiveness values).
- We tended to retain options with low variability in SME ratings (indicating high agreement for effectiveness) and high variability in Soldier ratings.
- Options with negative or near-zero option-total correlations were flagged for review and possible deletion.
- A minimum of 20 items were retained on each test.

Item Selection Results for 31B

In terms of content overlap, none of the scenarios developed for the 31B SJT seemed similar enough to warrant their removal from the test. The general themes of the scenarios were similar for a few of the items; however, the options were distinct enough to retain those items. Two options were deleted from items due to high SME standard deviations (>2.00). Three items were deleted because low option-total correlations reduced the number of options to fewer than four. The final total number of items was 24.

Item Selection Results for 91W

As with the 31B SJT, no items on the 91W SJT needed to be removed due to redundant scenario content. One item was deleted because of restricted distance (i.e., <2.0) between the highest and lowest SME ratings of effectiveness. Another item was deleted because it had only three remaining options after one of its options was dropped; the dropped option had an SME standard deviation above 2.00. At this stage, the test had 22 remaining items. An item with only four options contained an option with a negative option-total correlation. Thus, the option—and the item—were dropped. The final 91W SJT form had 21 items.

Score Computation

For the MOS SJTs, a separate score was computed for each option using the Soldier's effectiveness rating of the option. The option score was the distance between the Soldier's rating and the option's keyed effectiveness. Using this algorithm, lower scores are better. Because we wanted higher scores to indicate better performance, we reversed the scale by subtracting it from six. The final algorithm is shown in formula 1 below:

$$\text{Option Score} = 6 - |\text{SME mean} - \text{Soldier rating}| \quad (1)$$

The total score was computed as the mean of all option scores. Thus, for the option scores, and total scores, the lowest possible score is zero and the highest possible score is six. In reality, though, the lowest possible score is slightly above 0 and the highest possible score is slightly below six because the keyed effectiveness values are rarely integers, whereas the Soldiers' ratings are always integers. For example, if an option's keyed effectiveness is 4.5 then the closest a Soldier's rating can get to the key is 0.5 (i.e., with a rating of 4 or 5).

Descriptive Statistics and Reliability Estimates

Overall Sample

Soldiers who did not complete at least 70% of the test were screened out prior to analyses. For the LeadEx, 22 cases were removed. For each MOS-specific test, three such cases were removed. Table 5.1 displays the descriptive statistics and internal consistency reliability estimates for the final test scores. All scores show reasonably high levels of reliability and sufficient score variability.

Table 5.1. Descriptive Statistics and Reliability Estimates for SJT Scores

Composite Score	n^a	*M*	*SD*	Minimum to Maximum (Range)	Coefficient alpha
LeadEx	624	74.79	12.54	34.76 – 93.90 (59.14)	.82
31B	111	4.58	0.30	3.11 – 5.03 (1.92)	.90
91W	135	4.74	0.27	3.80 - 5.12 (1.32)	.80

Note. The LeadEx score is the percentage of possible points earned. For the MOS SJTs, the score can range from zero to six. Sample sizes for the MOS SJTs are small for coefficient alpha due to listwise deletion (31B $n = 38$, 91W $n = 55$ for Soldier ratings). The 91W test had 84 options and the 31B test had 96 options. All three SJTs had four options per item.

[a] Sample sizes here reflect the requirement that coefficient alpha computations use list-wise deletion for Soldiers with missing data.

Subgroup Analyses

LeadEx Subgroup Differences

For the LeadEx, we were able to conduct subgroup analyses by gender and for most race/ethnic groups because each subgroup had the required minimum of 20 cases. Table 5.2 shows the subgroup differences for the LeadEx SJT. For comparison, we have included results from two prior data collections using this form of the LeadEx – the NCO21 project concurrent validation (Waugh, 2004) and the PerformM21 Phase II data collection (Knapp & Campbell, 2006). As in previous data collections, female Soldiers scored somewhat better than male Soldiers and White Soldiers scored somewhat better than Black Soldiers. Although these findings are generally consistent with prior research on the LeadEx, the size of the effects varies considerably. The differences in effect sizes across the administrations are even more pronounced when comparing Asian-White and Hispanic-White subgroups, though this may be because of the relatively small non-referent group sizes. These differences might be caused by differences in the characteristics of the samples. The three samples differ considerably with regard to MOS mix and the relative mix of females and males within those MOS. They are also not completely comparable with regard to pay grade. It is quite possible that these factors influence observed effect sizes. There were no differences in deployed status (i.e., having been deployed in last 2 years or not) on the LeadEx.

Table 5.2. Subgroup Differences in the LeadEx Scores

	n	*M*	*SD*	Effect Size	*p*	Phase II Effect Size	NCO21 Effect Size
Gender							
Female	88	77.23	10.96	0.24	.02	.08	.40
Male	490	74.13	12.96				
Race[a]							
Hispanic	56	75.28	9.96	.00	.81	-.45	n/a
White	295	74.84	13.06				
Asian	20	76.06	12.55	.07	.78	-.28	n/a
White	422	75.21	12.71				
Black	84	71.33	12.69	-.30	.03	-.46	-.26
White	422	75.21	12.71				
Black	84	71.33	12.69	-.50	.02	n/a	n/a
Hispanic	71	76.35	10.01				

Note. Effect sizes are calculated as (Mean of non-referent group – Mean of referent group)/*SD* referent group. Referent groups are Male and White for gender and race, respectively.

[a] Soldiers were allowed to select more than one race/ethnicity which resulted in varying sample sizes. In instances where a Soldier was a member of both groups (i.e., White and Black), he was assigned to the minority group for analyses.

31B Subgroup Differences

Table 5.3 displays the subgroup differences for the 31B SJT scores. There were no subgroup differences by gender, but the sample size for females was small, suggesting this finding should be interpreted with caution. In terms of race, there were not enough Black Soldiers to assess White/Black differences, so race differences were compared as White/non-White. White Soldiers scored somewhat higher than non-White Soldiers, but the sample size for non-Whites was quite small. As with the female group, this finding should be interpreted with caution.

Table 5.3. Subgroup Differences in the 31B SJT Scores

	n	*M*	*SD*	Effect Size	*p*
Gender					
Female	21	4.54	0.29	-0.13	.59
Male	90	4.58	0.30		
Race					
Non-White	20	4.45	0.29	-0.52	.04
White	91	4.60	0.29		

Note. Effect sizes are calculated as (Mean of non-referent group – Mean of referent group)/*SD* referent group. Referent groups are Male and White for gender and race, respectively.

91W Subgroup Differences

For the 91W SJT scores, subgroup differences were assessed for gender and race (White/Black) differences, but the sample sizes for the non-referent groups (i.e., females, Blacks) were still quite small. Thus, results of subgroup differences should be interpreted with caution. Table 5.4 shows that there were no subgroup differences by gender. However, there was a significant difference in 91W SJT scores for race, such that White Soldiers scored higher than Black Soldiers by an appreciable margin. Again, however, the number of Black Soldiers in this sample was quite small.

Table 5.4. Subgroup Differences in the 91W SJT Scores

	n	*M*	*SD*	Effect Size	*p*
Gender					
Female	27	4.75	0.21	0.06	.77
Male	108	4.74	0.28		
Race					
Black	26	4.53	0.20	-1.23	.01
White	97	4.78	0.39		

Note. Effect sizes are calculated as (Mean of non-referent group – Mean of referent group)/*SD* referent group. Referent groups are Male and White for gender and race, respectively.

Correlations Between Army-Wide and MOS-Specific SJT Scores

We computed the correlation between the LeadEx and each of the MOS-specific SJT scores. The correlation with the LeadEx was .20 ($n = 109, p = .03$) for the 31B SJT and .29 ($n = 127, p < .01$) for the 91W SJT. These correlations are low to moderate in size, suggesting

that the two types of SJTs are tapping sufficiently different content to justify using both measurement methods (i.e., Army-wide and MOS-specific). It is possible, however, that the correlations are attenuated by the different response methods used in the Army-wide and MOS-specific instruments. In an operational program, we would recommend using the same method (i.e., the effectiveness rating process used on the MOS-specific tests) for all SJTs. This is discussed further below.

Soldier Reactions to SJTs

Soldiers were asked how well they thought they did on the SJTs. As shown in Table 5.5, 31B Soldiers think they did better than 91W Soldiers on the MOS SJT. However, 91W Soldiers think they did better than the 31B Soldiers on the LeadEx. Also looking within-MOS, 31B Soldiers rated their performance on the MOS SJT higher than their performance on the LeadEx. This pattern was reversed for the 91W Soldiers.

The SJTs in general were well-received. Soldiers preferred the format of the LeadEx (i.e., "Most/Least" selections) to the format of the MOS SJTs (i.e., effectiveness ratings for all options). At least one Soldier commented that he has found himself in situations similar to those in the LeadEx.

Table 5.5 Soldier Self-Assessed SJT Performance

	Very Well	Well	Neither Well nor Poorly	Poorly	Very Poorly
All Soldiers LeadEx (n = 642)	12%	42%	37%	5%	5%
31B Soldiers (n = 132)					
31B MOS SJT	15%	63%	19%	2%	1%
31B LeadEx	7%	39%	43%	6%	5%
91W Soldiers (n = 113)					
91W MOS SJT	8%	50%	33%	7%	2%
91W LeadEx	16%	47%	32%	3%	2%

Discussion

The SJT measurement method has been well-received by Soldiers and other Army personnel. Army research using SJTs similar to those described here indicates that the method yields useful criterion information for selection and classification research (Knapp et al., 2005), which casts a favorable light with regard to their use for routine performance measurement. Moreover, the evidence thus far indicates that there is value, at least for some MOS, to the inclusion of both Army-wide and MOS-specific SJTs.

The LeadEx subgroup score difference findings vary across the samples of Soldiers. None of the differences are particularly large (the largest effect size in Table 5.2 is -.50), so we do not believe such subgroup performance differences should negatively impact the value of this measurement method. We do, however, think it would be interesting to explore the data further to understand the fluctuations in findings across samples.

In the Phase II report, we suggested that operational SJTs use the "most and least effective" response format like that on the LeadEx. Subsequent research, primarily in the context of the Select21 project (Knapp et al., 2005), leads us to change this recommendation in favor of the effectiveness rating format like that on the two prototype MOS SJTs. Despite the preference of some Soldiers, there are several reasons for this change. First, the traditional strategy for scoring effectiveness rating SJTs involves a comparison of examinee effectiveness ratings to the mean SME rating. Thus, respondents can improve their scores simply by rating items in the middle of the 7-point scale (Cullen, Sackett, & Lievens, 2004). We have adopted a scoring strategy, however, that combats this weakness (see Waugh & Russell, 2005, for a detailed explanation). Another seeming advantage of the most/least response format is that it might take respondents less time to complete a test item. The Select21 research has shown, however, that the effectiveness rating format yields more reliable score information. This suggests that a test using the effectiveness rating format could have high reliability with fewer test items than a test using the most/least response format. Finally, the effectiveness rating format must be used during item pilot testing. Using the same format for experimental and operational test items will make it easier to embed new items into an operational test to collect the necessary pilot data.

Finally, we used a relatively unsystematic strategy for determining what types of content to include in the MOS-specific SJT scenarios, relying largely on a small group of SMEs to make this determination. We recommend following a fairly traditional critical incident analysis process when developing an operational SJT (Flanagan, 1954). Once the applicable dimensions (i.e., constructs) are identified, they are unlikely to change very much over time. There will, however, be a continuing need for fresh test item content. To help ensure that this content is relevant, it would be best to collect scenarios and response options directly from Soldiers. This could be a burdensome activity, unless it could be embedded in related training or Soldiers' development activities. Such strategies should be explored to help ensure maintenance of job-relevant, effective SJTs. A related issue is the development of alternate test forms. SJTs are notoriously multi-dimensional, and there is little research that suggests effective strategies for creating multiple test forms that are truly equivalent in terms of content and difficulty. Such research is necessary to support operational implementation of such tests in the Army.

CHAPTER 6: SIMULATIONS

Lee Ann Wadsworth (JPS, Inc)
Masayu Ramli, Chad Van Idekkinge, and Carrie Byrum (HumRRO)

Introduction

Computer-based simulations hold the potential for assessing Soldiers in a manner that closely resembles on-the-job demands without the complications of traditional hands-on work sample testing. We explored this concept using three distinct strategies:

- Development of inexpensive "low fidelity" simulations
- Development of a higher fidelity, complex simulation
- Adaptation of a training simulator for assessment

We developed three fairly simple simulation-based problems for the Armor Crewman (19K) Soldiers that were appended to their job knowledge test. Two problems used a multiple-choice response format and the third was a single-path simulation. All three problems related to machine guns.

The Patriot Air Defense Control Operator/Maintainer (14E) MOS was selected for assessment using a computer simulation because of the numerous technology features associated with this occupation. Resource constraints limited us to development of a simulation for a single activity. Specifically, this complex (i.e., multiple-path), fairly realistic simulation evaluates how well Soldiers can resolve an azimuth fault at the radar set by following procedures. The simulation was designed to balance realism, affordability, and technical requirements.

One of the goals of the PerformM21 research was to explore the possibility of using existing technology for competency based testing. With this in mind, we attempted to adapt the Engagement Skills Trainer (EST) 2000, a training simulator used throughout the Army, for testing within the Military Police (31B) MOS.

The remainder of this chapter is organized into three sections, corresponding to the three approaches to simulation testing we explored. Note that the 14E azimuth fault simulation and the 31B EST 2000 simulation required data collection procedures that were different than those used for the other PerformM21 Phase III pilot tests. Therefore, the discussions provide additional detail about those data collections.

Low Fidelity Simulations

As mentioned, we constructed three "items" related to machine guns that were administered along with the 19K job knowledge test. The graphics used in these items were adapted from training programs used at the schoolhouse. The first simulation comprised four multiple-choice questions related to a .50 caliber machine gun that unexpectedly stops firing. After responding to each question, the correct response option was illustrated with ShockWave Flash animation and accompanying audio. For example, the first multiple-choice question asked "The gun fired 15 rounds and then quit firing. What should the TC announce?" and displayed

four response options: "Jammed," "Fire," "Misfire," and "Stoppage." After the Soldier marked the box next to his choice of the correct response option, he clicked on a box marked "Submit Answer." This was followed by an animation that showed a tank commander firing a .50 machine gun. The Soldier heard the sound of the machine gun firing. After firing 15 rounds, the machine gun stopped firing and the Soldier heard the tank commander say the correct response, "Stoppage." Soldiers were given one point for each correct answer to the four component multiple-choice questions.

The animation and audio used in this simulation had been developed previously for a training application, which is why its focus is on increasing the Soldier's understanding of the correct response to each of the four questions. This had the disadvantage of not really helping the Soldier understand the question, itself. It had the advantage of making sure the Soldier was aware of the correct response to each question as he progressed through the simulation.

The second simulation was a single multiple-choice question illustrated by a ShockWave Flash animation. In this simulation, the animation began as soon as the question was displayed and required no input from the Soldier before commencing. The animation displayed the top view of an M2 machine gun and revealed that the GO end of the M2's headspace gauge would not enter the headspace. The text of the multiple-choice item directly related to the animation by describing the animation content and asking for the appropriate next step given the situation displayed in the animation (and described in the multiple-choice item stem). In this way the animation supplemented the Soldier's interpretation of the multiple-choice question. The item was scored by allotting a single point to the correct response.

Finally, the third simulation was a ShockWave Flash animation that guided the Soldier through a sequence of 11 steps involved in performing a function check on the M2 machine gun. The Soldier was presented with a graphic of the machine gun, informed that the headspace and timing were set on the gun, and instructed to proceed with a function check of the gun by clicking on the appropriate location on the machine gun for each subsequent action. Therefore, in order to progress through the simulation, the Soldier was required to select the next action to be performed. After the Soldier clicked on a machine gun location, feedback was provided that indicated whether his selection was correct or incorrect. Furthermore, irrespective of whether the action location selected was right, both text-based feedback and animated sequences revealing the correct step in the function check were displayed. Each step in the sequence was scored one point if done correctly, for a total of 11 possible points.

Although we did not include any survey questions to ask Soldiers what they thought about these three low fidelity simulations, they expressed considerable enthusiasm for the items during testing and in the focus groups that followed. Specifically, Soldiers' preference for items grew concomitantly with the integration of animation into the context of the question.

We made use of pre-existing animation programming as a cost-saving measure. This strategy, however, limited the content of what we could assess and made it hard to take full advantage of animation to illustrate the entire problem. Moreover, it was still time-consuming to adapt prior programming to the PerformM21 test environment. In an operational situation, we would advise looking for available animation to support test development, but would discourage trying to force fit what is available into what is needed. On the positive side, it was easy to score

these simulation-based items, since the first two used a multiple-choice response format and the third was a single path simulation with very distinct scoreable steps.

Azimuth Fault Simulation

Description of Test and Supporting Materials

The computer-based prototype azimuth fault simulation is a scenario to evaluate whether 14E Soldiers can resolve an azimuth fault at the radar set by following procedures (either with or without using their technical manuals). The simulation incorporates the ability to operate equipment and communicate, including audio of other team members and the section chief. Soldiers can move through the scenario and manipulate equipment using the computer mouse. Although many pieces of equipment appear to react when pressed, there are only two primary active paths available to fix the azimuth fault, with each providing multiple optional steps a Soldier may take. With further development, all reasonably possible paths could be programmed. Two multiple-choice questions pop up at key points in the simulation. There are also audio prompts providing realistic team communication as well as stress when the Soldier takes an incorrect action. Additional descriptive detail is provided in the later section on score development.

Because Soldiers have differing levels of experience with computers and the environment we created, we developed a Quick Start Guide (QSG) to familiarize them with how to navigate and operate the equipment in the simulation. In addition, since the mechanics of administering the simulation are different from the other PerformM21 pilot tests, we also developed the *14E Supplement to the Phase III Test Administration Manual* (14E TA Manual).

Quick Start Guide

The simulation is a self-contained module that enables Soldiers to navigate within and around the Engagement Control Station (the van) and the radar set, operate equipment in various panels, and access the Interactive Electronic Technical Manual (IETM). While the simulation has been described by the SMEs and Soldiers as having a high level of realism, working through the simulation is not the same as operating the actual equipment. Therefore, we developed the QSG as a self-paced, interactive guide to help familiarize Soldiers with how to operate within the simulation before taking the actual test.

The QSG includes sections on starting the simulation, navigation, interaction with the equipment (e.g., opening doors/panels and turning off/on switch indicators), using the manuals, and communication. For most Soldier actions, the QSG provides feedback such that if the action is not correct, the QSG gives hints on how to perform the proper action. Figure 6.1 presents a screen capture of one of the pages in the QSG.

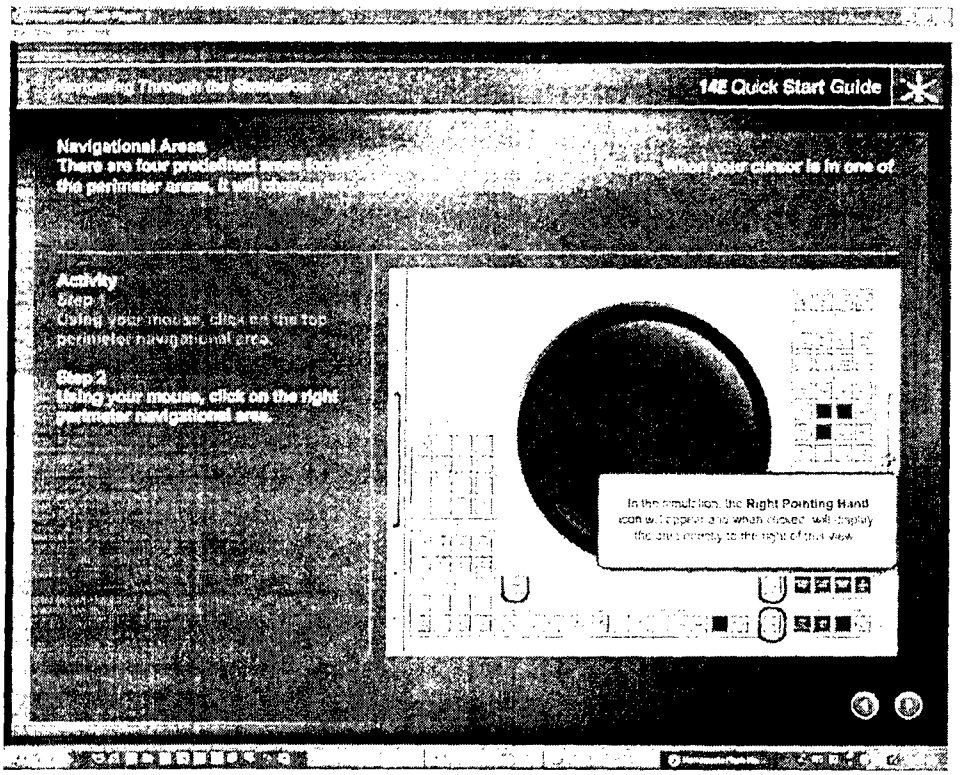

Figure 6.1. Screen shot from Quick Start Guide.

Pilot Test Data Collection

Test Administration

The one 14E data collection occurred at Fort Bliss, TX. The test sessions were divided into two parts, each of which lasted approximately one hour. During the first part, the Soldiers took the prototype simulation test and received feedback on their scores. In the second hour the Soldiers logged into the server to complete demographic and simulation-related questions and then completed the Army-wide and LeadEx test items. Both during and after the testing sessions, Soldiers participated in informal, short focus groups/interviews about their opinions of the simulation, job knowledge test items, and the testing process in general.

Technical Problems

We experienced problems with technology both in the simulation and with the server. Those involving the server were described in Chapter 3. The simulation-based problem is described in this section.

While designing the simulation, the SMEs suggested that all switch indicators appear as though they work, even though they would not necessarily function according to expectation because of the limited nature of this prototype simulation. For example, at Man Station One in the van, if an incorrect switch indicator was activated on the console panel, it would visually appear to work, but it would not bring up the expected display on the computer screen. Anticipating this, the

script in the 14E TA Manual provided explicit instructions describing that because this was a prototype test, not all switch indicators would work and that Soldiers should use only the specific equipment required to resolve the fault and not operate/open additional switches/panels. This design decision resulted in two unanticipated issues that occurred while the test was administered.

First, when administering the prototype test, many of the Soldiers tried relevant but not essential switch indicators to get more information about the fault, causing them to lose track of the last thing they did correctly. The only way to bring the Soldier back to the point of departure was to review the log files, which were not accessible while the Soldier was taking the simulation. Second, some Soldiers tried to explore the flexibility of the simulation during testing even though they were discouraged from doing this and were provided ample opportunity to try out the equipment with the QSG. Nonetheless, some were curious about the possibilities and flexibility of the simulation and pushed many more switch indicators than otherwise would have been reasonable, inflating the number of switch indicators and potentially also inflating the number of panels accessed. This was particularly evident in the high number of incorrect switch indicators pressed (highest number was 256) and incorrect panels accessed (highest number was 15). Both of these actions, or errors, were used in creating scores.

Both of these activities (accessing the log files or pressing excessive switches) caused the simulation to freeze and show a white screen, at which point the only option was to exit the simulation before finishing. If time allowed, Soldiers were allowed to restart the exercise. In an operational testing situation, it may be appropriate to consider providing Soldiers time to "play" in the simulated environment before starting the official test. It also would be helpful to provide access to the log files without having to exit the simulation. The existence of these two problem areas undoubtedly affected some of the scoring outcomes in the pilot.

Scoring the Simulation

Development of the scoring plan for this simulation was challenging. With multi-path simulations, there are often many "opportunities" to earn points. Decisions must be made concerning what those opportunities are and the value of those points. The SME group opted to set the total base points at 100. From there, they wanted to award bonus points for knowing certain shortcuts and penalties for taking certain unnecessary steps. Detailed scoring decisions are described below. The SMEs suggested that scores should be based on the following:

- Correctly following procedures.
- Successfully resolving the fault without calling maintenance.
- Accessing a minimum number of incorrect panels, switch indicators, circuit breakers, and toggle switches.

The SMEs also described Soldier errors that could result in fratricide and/or very expensive equipment repair or replacement. While not incorporated in the design of the current prototype, we agreed with the SMEs that it would be appropriate to include penalties for such errors in future modifications if this scenario is taken to the next step in development. Points were awarded as follows:

- *Basic procedures in the van (40 points).* Soldiers received these points by being able to reach the second multiple-choice question. The Soldier's ability to follow standard procedures to reach this point is an indication of minimal skills in the van. If Soldiers could not get this far, their score would be 0, indicating they did not know how to use the equipment or their manual.

- *Number of multiple-choice questions marked correct (20 points total).* There were two multiple-choice questions.

    - *Fix or fight question (10 points).* This was an embedded knowledge question where Soldiers were required to indicate whether the radar set azimuth fault should be fixed or they should continue fighting. The correct response is that they must fix the fault.

    - *Call maintenance or troubleshoot (10 points).* This question evaluated whether Soldiers follow procedures requiring that the Tactical Control Assistant (TCA; the role the Soldier is playing in the simulation) notify maintenance (the chief) about any problems. There is a written communication sequence wherein the chief asks if the TCA wants to request maintenance help or continue troubleshooting. According to our SMEs, the proper procedure is to continue to troubleshoot and try to resolve the problem without additional help. If Soldiers opted to request maintenance help, they were exited from the simulation.

- *Resolved the circuit breaker (10 points).* The Soldier earned these points by finding and fixing the circuit breaker that caused the fault.

- *The Soldier resolved the fault (30 points).* The Soldier earned these points by finishing the simulation, and taking all final steps needed to correctly resolve the azimuth fault.

- *The Soldier was awarded points for knowing short-cuts (up to 10 points).* The standard path takes 22 steps to resolve the fault. Soldiers were not penalized for taking additional steps as long as they did not involve accessing incorrect panels/switches. Knowledge of short-cuts could result in Soldiers correctly resolving the fault in as few as 12 steps. Correctly completing the scenario in fewer than 23 steps resulted in the Soldier being awarded points. Starting with the minimum 12 steps, 10 points were awarded, decreasing by 1 point for each additional step. For example, if Soldiers completed the scenario in 15 steps they were awarded 7 short-cut points, and if they completed the scenario in 23 steps, they were awarded no short-cut points.

There were potential deductions for two types of incorrect actions.

- *Deduct points for number of incorrect panels accessed.* For each incorrect panel, we deducted 1 point. Lower numbers in this category demonstrated higher levels of Soldier expertise. The range of incorrect panels accessed was 1-15.

- *Deduct points for number of incorrect switches used.* As described earlier, Soldiers pressed many more incorrect switch indicators than panels (the range was 1-256). Yet pressing incorrect switch indicators is a mistake similar to that of incorrect panels. Since the range of incorrect panels accessed was 1-15, to keep the points deducted for incorrect switches on the same scale, we divided the number of incorrect switch indicators by 10 and deducted the result from the Soldier's score. The range of points deducted was .1-25.6. As with the incorrect panels, the lower the number of incorrect switches used, the higher the expected expertise of the Soldier.

Two additional items were tracked by the simulation but did not impact their scores - one because of duplication, and the other because the test was not designed to be timed.

- *Number of audio prompts.* Each time Soldiers accessed an incorrect panel or used an incorrect switch, they received an audio prompt telling them they took an incorrect action (e.g., "Do you need help?" or "Chief wants to know what's taking so long."). These audio prompts, designed to increase the stress for the test taker, duplicate the deductions above and are therefore not included in the scoring.

- *Amount of time in minutes to complete the scenario.* While we tracked the Soldier's time to complete the scenario, this is not appropriate to include in the scoring since the Soldiers were not told that the test was timed. However, it is expected that the time to complete the simulation and the number of steps to resolve the fault would be highly correlated.

Although the goal was a scoring system with a maximum of 100 points, bonus and penalty points resulted in an actual range of possible scores from -40.6 to 110. We felt that conceptually this was difficult to explain and might detract from the results we were reporting. To rectify this, we recoded the scores so that the lowest possible score was 0 and the highest was 150.60. It is important to note that this transformation does not affect any analyses conducted.

Pilot Test Score Results

The observed scores ranged from 24.20 to 150.00 with a mean of 95.11 and a standard deviation of 31.43. Because the 14E sample included only 69 Soldiers, of which 68% indicated they were both White and non-Hispanic, most meaningful subgroup comparisons are not possible. We were able to compare the scores of Whites versus non-Whites, however, and found that while White Soldiers scored somewhat higher than non-White Soldiers, the difference was not statistically significant (see Table 6.1).

Table 6.1. 14E Simulation Score Subgroup Differences

	n	*M*	*SD*	*d*
White	39	100.25	31.94	.37
Non-White	30	88.42	29.97	

Note. Effect size was calculated as the difference between the means of the two groups divided by the standard deviation of the White group.

We collected information from Soldiers about their use of computer games to determine whether this had any affect on their performance in the simulation. Approximately 66% indicated that they do play computer games. Of these, 52% said they played less than 20 hours per week, while 48% indicated they played 21 or more hours per week. As for game variety, 38% said they play one to two games, 31% indicated they played three to five games, and 31% indicated they played six or more games per week. Table 6.2 shows the results of the analyses. None of these group differences were significant. We also looked at performance based on deployment status. Earlier we noted that 31% were deployed OCONUS within the previous 2 years (see Table 3.3). The differences between those who had been deployed ($M = 95.92$, $SD = 31.30$) and those who had not ($M = 95.40$, $SD = 35.21$) were negligible.

Table 6.2. Mean 14E Simulation Scores by Soldier Computer Gaming Experience

	n	*M*	*SD*	*d*
Do You Play Computer Games?				
Yes	45	97.44	33.81	.21
No	24	90.74	26.53	
How Many Hours per Week Do You Play?				
1 to 20	24	96.14	29.54	.08
21 or More	21	98.93	38.82	
How Many Different Games Do You Play?				
1 to 2	16	97.61	31.14	.06 - .23[a]
3 to 5	14	103.26	33.80	
6 or More	14	95.48	36.32	

Note. Effect sizes were calculated as the differences between the means of the two groups divided by the pooled standard deviation.

[a] The effect size differs with the exact comparison being made (1 to 2 vs. 3 to 5 = .16; 1 to 2 vs. 6 or more = .06; 3 to 5 vs. 6 or more = .23).

Soldier Reactions

Focus Groups

The Soldiers' reactions to the simulation were generally positive. Many Soldiers said they did not need the QSG because they were so familiar with computers, while a few said it was helpful. The Soldiers thought the realism and interactivity (e.g. navigation to different equipment, use of manuals) was good. However, the Soldiers generally felt too limited with only two paths to complete the simulation. They said if they could have had more options they could have fixed the problem right away. Some of the Soldiers said they had difficulty understanding the feedback of the "Hard Copy" switch indicator, and they sometimes were concentrating so much that they missed an important audio prompt and wished they could have had a mechanism to hear it again.

Technical problems did arise, largely because the Soldiers did not stop right away when they heard the "Do you need help?" audio prompt alerting them to an error. They got frustrated when they constantly heard that prompt. Soldiers said they would like to have received hints from the simulation about what to do when they got stuck.

Most Soldiers indicated that they liked having a simulation for assessment purposes, and some said they liked it better than paper-and-pencil because they have to know the job in order to resolve the problem, and when on the job, they cannot rely on the study guide. But others thought it was "too fake" and "too textbook" because they could not skip what they described as unneeded steps. Some also said they do it differently in the field and that the book paths are not always right.

In addition, there were some general comments that related to the Soldiers' experience as a 14E. Some said their chief would never let them touch the radar set so their "correct" answer to the second multiple-choice question was the wrong answer in the simulation (i.e., Call maintenance for help vs. Continue troubleshooting). Others said the simulation was not fair because they spend all of their time in the Information and Coordination Central (ICC) or the Antenna Mast Group (AMG) and never get in the van, so they had no idea what to do. When asked if they could follow the manuals, some said they were able to follow it while many said they tried but got so frustrated they stopped.

Survey Questions

Soldiers provided their opinions about the simulation. For the most part, this feedback was positive. Sixty-one percent either agreed or strongly agreed that the level of realism was acceptable. This is especially encouraging because, as previously mentioned, we were not able to make all the switch indicators functional and some Soldiers noted in the focus groups that they felt limited because the simulation allowed for only two primary paths to correct the fault. When asked if there is value in using simulations to evaluate 14E skills, 56% agreed or strongly agreed. Sixty-seven percent of the Soldiers tried to access the IETM (electronic manual), and of those roughly 77% found it easy to access and 47% were able to easily manage the simulation and the IETM together. Table 6.3 shows other feedback concerning the ease of progressing through the simulation. Note that for these three questions, between approximately 57% and 62% answered "easy" or "moderately easy." Again, this is a very positive sign given the limitations mentioned above, and the fact that approximately 51% exited early due to problems with the simulation (e.g., they got stuck or received the "white screen" noted above).

Table 6.3. 14E Soldier Feedback Concerning Ease of Use of Simulation

	Easy	Moderately Easy	Neither Easy nor Difficult	Moderately Difficult	Difficult
Operating Equipment in the engagement control station	28.6%	28.6%	21.4%	15.7%	5.7%
Operating the radar set	27.1%	32.9%	27.1%	5.7%	7.1%
Navigating Around the Simulation	38.6%	22.9%	30%	2.9%	5.7%

Discussion

While the azimuth fault simulation has been well-received by those Army personnel who have seen it and by most of the Soldiers who took the test, there were a number of problems associated with it. Most of the problems we experienced were due to limited development resources and lack of Soldier preparation, both of which could be more fully addressed in an operational testing program.

As discussed, the limited number of paths that were programmed into the simulation, combined with the examinees' inexperience with the simulation environment, led to some Soldiers having problems navigating the simulation. These problems were exacerbated by the fact that some 14E Soldiers did not have much experience working in a van. This latter concern is not unique to the azimuth fault simulation – all of the PerformM21 tests include material that is not particularly pertinent to a Soldier's current assignment. Even so, the problems would have been alleviated by (a) programming additional paths into the scenario, (b) conducting more extensive beta testing prior to administration to actual examinees, and (c) allowing Soldiers more time to familiarize themselves with the simulation prior to testing. Indeed, it would be desirable in an operational program to have a practice version of the simulation available to Soldiers 24-7 on the Internet.

We had strong support from the 14E proponent point of contact and a core group of SMEs that was dedicated to the project from the time we identified the scenario through final production and initial testing. In an operational setting, however, we suggest having a core group, but also rotating SMEs from different units with different specialties through the process. Some Soldiers noted that unit Standard Operating Procedures (SOPs) require them to request maintenance support for faults at the radar set. Recall that such action in the simulation caused the simulation to end and the Soldier to receive minimal points. Some even said that we had the compulsory steps at the radar set wrong, which our SMEs had previously verified as correct. Had we been able to involve a broader sampling of SMEs in the development process, we may have learned about differing SOPs with regard to fixing faults in the radar set or calling maintenance.

One of the bigger issues with respect to developing a realistic job simulation such as this is the up-front costs. This project has demonstrated that while the costs to develop a simulation are high relative to traditional knowledge tests, they can be mitigated in several ways. First, the test can incorporate a variety of methodologies and items (e.g., multiple choice, situational judgment test items, simple visuals and/or audio/video clips) as well as a simulation, which can range from the simplistic "user as passive recipient of information" to a more complex interface with the user making a variety of responses to different cues.

Second, to the extent that the equipment developed in one simulation for one skill level of an MOS can be used for other skill levels, the incremental cost to modify the scenario is relatively less, and the up-front cost to develop the equipment can be spread across multiple tests. Third, we believe that there are reusable software components from this effort that can be used to streamline the development of simulations for other MOS. For example, in this project we developed a process model that has the potential to underlie simulations supporting a broad range of MOS. Further work should be done in defining common assumptions that can be reused across most MOS, exploring visual editors to support automating the development of process flow charts, and enhancing the underlying architecture.

In summary, this experience provides support for the notion that a test of this type has value as a performance test for evaluating those kinds of skills and abilities that otherwise might require much more expensive procedures such as high fidelity hands-on testing. While the up front development costs of the initial architecture and equipment are high, each project thereafter (both within and across MOS) should require relatively less design and development work. Within one MOS, the more projects using the same equipment and thus requiring relatively minimal design and development work, the more the cost can be spread across multiple uses and the lower the per project costs. Further work in developing this process model and reusable architecture could have applications for testing both within and across MOS, as well as for self-guided assessments.

The Engagement Skills Trainer (EST) 2000 Assessment

Description of Test

The EST 2000 is a virtual weapons training system developed for the Army. The EST simulator provides marksmanship, collective, and judgmental shoot-don't shoot (SDS) training. The EST simulator is equipped with a wide screen, simulated weapons, and standard EST equipment requirements such as a projector, compressor, speakers, and telephone lines. The EST room is kept dark when the simulator is in use so that images on screen can be seen clearly. The marksmanship component provides various simulations, such as a military police qualification course, in which objective data, including weapon angle, trigger pressure, and shooting accuracy can be obtained.

The SDS component includes video-based situations that require Soldiers to interact with characters and determine if and when to shoot. There are 15 MP-specific SDS scenarios, each of which lasts about 1-2 minutes and includes 8-12 possible outcomes that trainers control based on how Soldiers handle the situation. Unfortunately, there is no objective way to measure performance on the SDS scenarios. Rather, trainers provide Soldiers verbal, qualitative feedback at the end of each scenario. Given this, we developed rating scales with behavioral anchors that will allow for a more objective evaluation of performance on this component of the EST.

We worked with SMEs, (i.e., primarily 31B instructors who operate the EST) to determine (a) the performance dimensions the SDS scenarios could be used to evaluate and (b) observable behaviors that exemplify low, moderate, and high performance on each dimension. We identified five dimensions that appear to be measurable with the SDS scenarios: communication skills, judgment, reaction time, marksmanship, and technique.

Pilot Test Data Collection

Training

We pilot tested the EST marksmanship and SDS components in one data collection at Fort Leonard Wood, Missouri. We began by training four Basic NCO Course (BNCOC) instructors how to administer the EST. Instructors were E6/E7 NCOs in the 31B MOS.

Three of the four instructors were trained to observe and rate Soldiers on the SDS component of the EST. Rater training was conducted in two parts. In the first part, we discussed the two EST components that would be administered and how examinees' performance on each component would be scored. Specifically, the simulated MP qualification course would be used to assess performance on the marksmanship component, and five MP-related scenarios would be administered for the SDS component. Based on length, content, and quality, we selected the following SDS scenarios: Domestic Disturbance, Armed Forces Bank Robbery, Shoppette Robbery, Electronic Store Robbery, and Felony Traffic Stop. Also, an EST M9 pistol (with the same heft and recoil as the standard MP sidearm, but modified to interact with the simulator and track firing speed and accuracy)would be provided to the examinees for both components. Then, we gave raters an overview of the rating scales that we developed. We discussed the competencies the scales were intended to assess, the various levels of performance they described, and specific behaviors under those levels of performance. Finally, we offered raters suggestions for using the rating scales and tips for avoiding common rating errors (e.g., leniency error, recency effect).

In the second part, we asked instructors to simulate the pilot test by performing as examinees would during the test. One instructor operated the EST while the other three served as raters. Using one of the raters as a shooter, we ran through both portions of the EST assessment. At the end of the SDS scenarios, the remaining two instructors evaluated the shooter's performance using the rating scales. We continued practicing the SDS portion in this fashion, where one rater served as a shooter, until all raters had practiced rating twice.

The fourth instructor was trained to follow specific instructions for operating the simulator. As mentioned, the SDS scenarios offer 8-12 possible outcomes that the operator controls. Therefore, to standardize the assessment, we developed a set sequence for each of the five scenarios. We gave the operator specific instructions for escalating/de-escalating the sequences at specific points in the scenario. We also advised him how to handle mistakes in the event he inadvertently failed to escalate/de-escalate in time.

Additionally, we instructed the operator to record the number of hits and misses from the SDS component. The simulator actually provides separate scores for misses and lethal and non-lethal hits. However, the instructors informed us that even though the simulator separates lethal and non-lethal hits, the distinction between the two is not that important. What matters most in the field is whether the MP hits or misses the target. Therefore, we combined the number of hits regardless of type. Also, we instructed the operator to run the simulator for the marksmanship component and to record examinees' performance data from the qualification course. In the second half of training, the operator was given an opportunity to practice these tasks.

Test Administration

Examinees. Eight Soldiers (7 males, 1 female) participated in the EST pilot test. All examinees were in the Active Component of the Army and in the 31B MOS. Most of them were White (there were two minority participants), in the E6 pay grade (notably more senior than our E4 target audience), and on average were 24 years old. We recorded the amount of EST experience each examinee had and found that approximately half of them had used the simulator in the past. Halfway through testing, we realized that some of those who had prior EST

experience did not have SDS experience. Given this, we started to record more specific EST (i.e., SDS) experiences.

Examination schedule/procedure. The EST pilot test took place in two locations. Participants reported to the administration room where they completed a brief (approximately 5 minutes) computer-based background form. Then, they reported to the second location, the EST Warrior Room. We scheduled one participant per hour for the pilot test. As examinees arrived, they signed in and completed the initial paperwork (i.e., the background information form and the Privacy Act Statement). The test administrator then escorted examinees to a room that contains one of the EST simulators. There, the test administrator reviewed the instructions for taking the assessment, which included a description of the measures to be administered, the weapon they will use, the level of force allowed, and how they should respond to the SDS scenarios.

After the short briefing, examinees were given a few minutes to practice handling the simulated M9. Practice time was important as it afforded examinees an opportunity to become familiar with the weapon (such as when locking and loading) and the way the system responded to certain actions. For example, re-holstering the weapon a certain way could cause the system to "lose" one round of ammunition. Next, the operator gave examinees feedback on their weapon handling skills. Then, the operator gave examinees 10 practice rounds on a 50m target. Following this, we began the six-table (i.e., one 7m table, two 15m tables, two 25m tables, and one 35m table) simulated MP qualification test. As practiced in training, the operator manually recorded examinees' hit rate as they completed the tables.

Next, we began the SDS portion of the simulation. As with the marksmanship component, we allowed examinees one practice round. We selected the practice SDS scenario based on its similarities to the evaluated scenarios. After the practice session, the operator began the evaluated portion of the SDS component. Recall that raters evaluated examinees' performance on five SDS scenarios. As each scenario played, raters observed the examinees' performance and took notes. The scenarios were replayed for raters so they could observe where shots were fired (e.g., in center mass of suspect's body). This process was repeated for each scenario. At the conclusion of the SDS component, the operator and raters provided examinees feedback on their performance. The testing process, from completing background information to receiving SDS performance feedback, lasted approximately 30 minutes.

Pilot Test Scores

Table 6.4 displays examinee scores on the EST-simulated MP marksmanship qualification course. Although there was little or no variation in examinee scores at the closer ranges of 7m and 15m, there was some variability at the farther ranges of 25m and 35m. This was particularly true of the 25m range, for which the "hit rate" ranged from 5% to 100%. Overall marksmanship scores ranged from 44% to 98%, with a mean of 83%.

Table 6.4. EST 2000 Military Police Marksmanship Qualification Course Scores

	Range								Overall	
	7m		15m		25m		35m			
Examinee	Hit/Shot	%	Hit/Shot	%	Hit/Shot	%	Hit/Shot	%	Hit/Shot	%
A	5/5	100	14/15	93	1/20	5	2/10	20	22/50	44
B	5/5	100	15/15	100	17/20	85	5/10	50	42/50	84
C	5/5	100	15/15	100	20/20	100	6/10	60	46/50	92
D	5/5	100	15/15	100	7/20	35	9/10	90	36/50	72
E	5/5	100	14/15	93	18/20	90	9/10	90	46/50	92
F	5/5	100	15/15	100	19/20	95	7/10	70	47/50	94
G	5/5	100	15/15	100	20/20	100	9/10	90	49/50	98
H	5/5	100	15/15	100	15/20	75	9/10	90	44/50	88
Mean	5/5	100	14.8/15	98.0	14.6/20	73.0	7.0/10	70.0	41.5/50	83.0

Table 6.5 presents descriptive statistics and correlations of ratings of examinee SDS performance. Several findings are noteworthy. First, the highest mean dimension rating was only 3.63 (out of a possible mean of 5.0), which indicates that these ratings did not suffer from a leniency effect often observed in ratings. In addition, there was decent variation in ratings both across dimensions and across examinees (mean dimension ratings ranged from 2.53 to 3.80 for the eight examinees). Although there was variation in mean ratings across dimensions, correlations among the dimensions (with the exception of marksmanship) were quite high.

Table 6.5. Descriptive Statistics, Correlations, and Reliability Estimates for EST 2000 Shoot-Don't Shoot Ratings

Dimension	*M*	*SD*	1	2	3	4	5	6
1. Communication	2.88	0.80	(.73)					
2. Judgment	3.08	0.97	.77*	(.70)				
3. Reaction Time	2.96	0.81	.78*	.78*	(.87)			
4. Marksmanship	3.46	0.59	.10	.06	.02	(.78)		
5. Technique	3.63	0.82	.72*	.87*	.84*	.14	(.50)	
6. Overall Mean	3.20	0.60	.88*	.91*	.90*	.28*	.92	(.82)

Note. $N = 8$. Statistics are based on ratings (on a 5-point scale) averaged across three judges. Interrater reliability estimates are intraclass correlation coefficients (C,3) and appear along the diagonal in parentheses.
*$p < .05$ (two-tailed).

Also shown in this table are interrater reliability estimates for the ratings of the three judges. Reliability estimates ranged from .50 for Technique to .87 for Reaction Time. The interrater reliability for the mean ratings was .82. These results suggest that judges rated the performance of the eight examinees in a similar way.

Examinee scores on the marksmanship aspect of the SDS are displayed in Table 6.6. Shown is the number of shots taken and targets hit in each scenario. Across scenarios, there was decent variation among examines in both shots taken (5 to 17 shots) and shooting accuracy (46.2% to 80.0%).

Table 6.6. EST 2000 Shoot-Don't Shoot Marksmanship Scores

	S2		S3		S4		S5		Overall		
Examinee	Shot	Hit	Shot	Hit	Shot	Hit	Shot	Hit	Shot	Hit	%
A	2	2	2	1	3	1	2	2	9	6	66.7
B	2	1	5	2	3	1	3	2	13	6	46.2
C	3	1	2	2	2	1	2	1	9	5	55.6
D	2	1	2	1	2	2	1	1	7	5	71.4
E	2	1	1	1	1	1	1	1	5	4	80.0
F	5	2	4	3	5	4	3	1	17	10	58.8
G	2	1	1	1	2	1	1	1	6	4	66.7
H	2	2	3	2	2	1	4	2	11	7	63.6
Mean	2.5	1.4	2.5	1.6	2.5	1.5	2.1	1.4	9.6	5.9	63.6

Note. S = scenario. Data for S1 are not shown because no examinees fired shots in that scenario.

The main finding was that the two EST components yielded moderate to high variability in scores across this small sample of Soldiers. Data also revealed that the estimated interrater reliability for SDS performance ratings was acceptable.

Soldier Reactions

In addition to gathering objective data, we asked instructors for general feedback on the use of the EST for competency based testing. Their feedback was overwhelmingly positive. For example, the instructors indicated that the rating scales we developed increased the usefulness of the SDS component. They felt that it would be useful to assess critical MP competencies before promoting Soldiers to the E5 level because communication skills, not just gross motor skills, are important at that level. Moreover, they commented that the rating scales adequately captured the intended dimensions. Of the five dimensions, they indicated that communication skills and judgment are the most important. However, they were concerned that MPs who have mostly combat experience (versus garrison law enforcement) would be at a disadvantage if this were to be used for promotion testing.

Discussion

The pilot test results indicated that in addition to training, the EST simulator has the potential to be used for competency assessment. Data from the marksmanship component suggested there was some variability in performance on the simulated MP qualification course and that this could therefore be used to discriminate high performers from lower performers. Results from the SDS component were also promising in that it showed decent variation in scores, and there was a high level of interrater reliability using the behaviorally oriented rating scales we developed. Even though initial results indicated that the EST has the potential to be used in competency based testing, several changes would have to be made to the instrument and the administration process before implementing this assessment.

Rating Scales

The first change is that the SDS rating scales would need to be revised based on instructors' comments and our observations during the pilot test. For the most part, these changes

are minor. In addition, results indicated that interrater reliability for the Technique dimension was low. Thus, this scale should be revisited in the future.

A more substantive change involves scoring the SDS marksmanship data. As stated, the operator recorded the marksmanship scores provided at the end of the SDS scenarios. We learned that raters used similar information to rate the SDS marksmanship scale (they observed the shot overlays shown during replays to determine where suspects were shot). Not surprisingly, results indicated there was a high correlation between the objective SDS marksmanship scores and the subjective marksmanship ratings. Given this, we suggest eliminating the marksmanship rating scale.

Participant Instructions

Examinee instructions. Recall that we gave examinees a brief overview of what to expect on the EST test, including how they should react to the situations presented to them. If an examinee did not sufficiently interact with the characters during the practice session, the test administrator and operator reminded him/her to do so for the evaluated portion. Even though much emphasis was placed on instructing examinees to respond to the scenarios as if they were real life situations, many of them did not react that way. Some of this could be due to the examinees' lack of experience with the simulator. For those who had not done the simulation before, speaking and reacting with characters on the screen may have seemed unnatural. They may have felt somewhat awkward to respond with the degree of realism we wanted. Another explanation could be that examinees were not motivated to take the exam. In an operational setting, perhaps examinees would react differently when they realize that they are in a high stakes situation and need to perform well. Regardless of the issue or method taken to correct it, additional emphasis should be placed on examinees responding realistically to the scenarios and be made clear that they are being evaluated on that aspect. We recommend providing several practice scenarios to give examinees time to become accustomed to the SDS component.

Not only did examinees have difficulty knowing *how* to respond to the scenarios, but they were also unsure of *when* to respond. For example, in the Domestic Disturbance scenario, some examinees were not sure if they should respond to the character that answered the door. In the Electronic Store Robbery scene, an examinee continuously gave commands to the suspect when it was not warranted (i.e., suspect did not display hostile intentions). Additionally, in the Felony Traffic Stop scenario, some examinees were uncertain of what to do at the beginning of the scenario when the getaway car remained idle for a few seconds.

It is possible that this confusion with how and when to respond represented training deficits, but it was project staff's impression that it was likely to due to unfamiliarity with the SDS simulator. Soldiers were only allowed one practice scenario, and the testing (i.e., scenarios) occurred quickly. Therefore, for operational testing, more detailed examinee instructions of what they could expect and various ways of responding should be provided. More specifically, instructions should indicate that sometimes they may need to interact with the screen right away and other times it may be more appropriate to wait for something to happen.

Operator instructions. Despite having extensive experience running the simulator, one operator error occurred during the pilot test in which the operator failed to de-escalate the

scenario at the appropriate time. During training, we instructed the operator to continue with the scenario if such a mistake occurred. However, this is not necessarily the best solution as continuing with the wrong sequence would result in the examinee receiving a different test scenario. The course of action the operator should take when he or she makes an escalate/de-escalate mistake in running the simulator needs to be determined and clearly outlined.

A better way to prevent operator escalate/de-escalate error is to standardize the process by programming the simulator to execute these decisions. We learned from EST developers that such capability (referred to as "free programming") exists in which the operator programs the sequence ahead of time, and it will automatically play the selected sequence during the exam. We did not use this option, because although the technology is available, additional funding would have to be spent to incorporate the change. For the pilot test, we decided it was not necessary to do this. However, if this instrument were to be used for promotion purposes, it would be worth investing in this feature.

Test Administration

Recording data. The operator was assigned the responsibility of recording scores from the marksmanship component. While this worked well for the pilot test, we recommend printing the data, if possible, to reduce operator burden and human error. The capability to print marksmanship data is available, but we did not use the printer during the pilot test because of a technical difficulty.

SDS practice scenarios. During the test administration, we realized that examinees needed more SDS practice time. About half of the examinees had not used the EST before and were quite unfamiliar with it. Although we allowed one scenario on which examinees could practice interacting and shooting, most of them (75%) missed the first shot on the first scenario that required use of deadly force. It also seemed that examinees began to interact with the characters more realistically with each scenario. By the second or third scenario, they issued more commands to suspects and spoke in an authoritative voice. Given this, we recommend incorporating more than one practice scenario for the SDS component.

Number of scenarios. During the instrument development phase, SMEs suggested that five scenarios were sufficient for assessing the SDS component. Indeed, results yielded acceptable interrater reliability by using just five scenarios. However, it may be beneficial to administer more scenarios to obtain a better reliability estimate. A repeated measures research design might be useful to determine the number of scenarios needed for good test-retest reliability.

Scenario replays. Raters requested that each scenario be replayed so that they could view the overlays of shot groups to rate examinees' marksmanship skills. At times however, raters disagreed with the simulator whether a shot was a hit or miss. That is, the simulator would register the shot as a miss, but during scenario replays, raters would say that the shot was a hit and adjusted the SDS marksmanship score accordingly. It needs to be decided whether the computer's judgment stands or if raters can override it. To keep the exam objective, it would be best to go with the simulator's assessment rather than allowing room for raters' interpretation, which could vary by rater.

Performance feedback. If the EST is used for operational testing, then the type of feedback examinees would receive must be determined. For instance, they could be given one overall score or feedback on each rating scale dimension.

Other considerations. If used for operational testing, new SDS scenarios would need to be created. EST developers estimate the cost of developing a new scenario is between $30,000 and $50,000. If the pilot test procedures were to be replicated, at least five scenarios would need to be developed. Furthermore, it needs to be determined if older versions of the simulator would be compatible with the new scenarios.

Finally, we only used law enforcement scenarios in the SDS component. SMEs mentioned that the 31B is moving towards combat-type duties and that it might be beneficial to include some combat-related scenarios to the SDS component. At present, approximately 15 such scenarios, called Infantry scenarios, are available. They present realistic situations 31B Soldiers face in combat.

CHAPTER 7: CROSS-METHOD RESULTS

Karen O. Moriarty and Deirdre J. Knapp

The goal of the MOS-specific portion of the PerformM21 project was to explore different testing methods, and one question of interest is how scores resulting from different methods correspond with each other. In this chapter, we briefly discuss the results of the cross-method analyses. It bears repeating, however, that with the exception of the 31B SJT and the 19K JKT, the prototype assessments administered in this phase underrepresent applicable test content. Therefore, these results should be interpreted with caution.

Table 7.1 shows the different tests that were administered to Soldiers in each MOS. Note that simulations were tried as well; however, except for the 14E MOS, those data are not included here. For the 31B MOS, only eight Soldiers took part in the simulation and, for the 19K MOS, only three simulation items were developed.

Table 7.1. Test Method by MOS

MOS	MOS JKT	MOS SJT	MOS Simulation	LeadEx	Common Core
Patriot Air Defense Control Operator/Maintainer (14E)			X	X	X
Armor Crewman (19K)	X			X	X
Military Police (31B)		X		X	X
Wheeled Vehicle Mechanic (63B)	X			X	X
Health Care Specialist (91W)	X	X		X	X
Other				X	X

The common core and MOS JKT correlations were previously reported in Chapter 4, and the LeadEx and MOS SJT correlations were reported in Chapter 5. Table 7.2 shows the cross-method correlations for each MOS. The common core score is based on the short form. For the most part, correlations are significant and moderate in effect size. The 63B results, however, are anomalous and difficult to explain. 63B LeadEx scores are uncorrelated with both the 63B JKT and the common core test scores. The relatively small sample size makes it difficult to conduct informative subgroup analyses (e.g., by test site) that might provide insight for these findings.

Table 7.2. Comparison of Cross-Method Correlations by MOS

Correlation	14E	19K	31B	63B	91W	Other
LeadEx & MOS JKT Scores		.26*		.00	.30*	
LeadEx & Common Core Scores		.54*	.28*	.09	.39*	.27*
MOS JKT & MOS SJT Scores					.27*	
Common Core & MOS SJT Scores			.47*		.26*	
MOS simulation & MOS JKT Scores	.19					

Note. 14E n = 31 ; 19K n = 41 – 59 ; 31B n = 93 – 109 ; 63B n = 56 – 69 ; 91W n = 106 – 126 ; Other n = 193

* $p. < .05$.

The other LeadEx results suggest that the LeadEx relationship with the common core short form is a little stronger than its relationships with the MOS JKTs. The LeadEx and common core short form correlations range from .09 to .54, and in the overall sample the

correlation is .37 (n = 519), p = .001. In fact, these correlations are equal to or greater than the correlations between the LeadEx and the two MOS SJTs. Recall from Chapter 5 that those correlations were .20 with the 31B SJT, and .29 with the 91W SJT, both of which are significant. It was noted that the correlations may be attenuated by the different response methods. If we standardize the MOS SJT scores within MOS and correlate them with the LeadEx, we get .27 (n = 236), p. = .001. One might expect the SJTs to correlate more with each other than with JKTs due to method effects (Millsap, 1998). However, these results do not support this notion. The 14E LeadEx and common core results are not reported because the sample size for that calculation is less than 20.

When looking at the common core test's relationships with the SJTs versus its relationships with the MOS JKTs, we find different results. From Chapter 4 we know that the correlations between the common core and the three MOS JKTs are .39 (91W), .40 (63B), and .72 (19K). The common core test's relationships with the SJTs range from .26 to .54. If we standardize the JKT scores within MOS and correlate them with the common core, we get .48 (n = 213), p. = .001. Similarly, if we standardize the MOS SJT scores and correlate them with the common core, the result is .37 (n = 204), p. = .001. This is not conclusive, but is suggestive of the expected method effects.

Summary

The obtained correlations here were not very large, suggesting that the different test types capture different portions of the Soldier performance domain. They appear to suggest an absence of method effects for SJTs in this population, but the presence of method effects for the JKTs. As we have noted, however, the results should be interpreted with caution. Not only do most of the tests suffer from overly-narrow content, but the sample sizes for some of these analyses are fairly small.

CHAPTER 8: SUMMARY AND RECOMMENDATIONS

Deirdre J. Knapp

Introduction

This report concludes the rapid prototyping and tryout portion of the PerformM21Army test program feasibility research effort. Through the course of this 3-year research program, we designed and developed prototype assessments targeted to E4 Soldiers seeking promotion to the E5 pay grade. We identified and developed a variety of assessment strategies (e.g., knowledge-based multiple-choice questions, situational judgment tests, simulations). These tests cover common core (Army-wide) content as well as content specific to five selected MOS. To the extent possible, we administered the assessments to Soldiers in a manner similar to that which would be used in an operational test program. Although we successfully used the Army's existing computer facilities for test administration and provided participating Soldiers with feedback on their test performance, the Soldier tasking process generally did not allow us to provide participating Soldiers with study guides to help them prepare for the tests in advance.

Our experience in Phase III of the PerformM21 research provided (a) lessons learned that can be used for planning an operational test program, (b) the basis for estimating some of the costs associated with an operational test program, and (c) tangible products (e.g., test items) that could be incorporated into an operational test program. In this final chapter, we summarize some of the lessons learned and the concrete products resulting from this research.

Lessons Learned

The Phase III pilot test experience demonstrated value for all of the measurement methods we tried. We were pleased by the Soldiers' reactions to the enhanced multiple-choice test method (i.e., the job knowledge tests), since this is the least expensive method to use, and provides for the most comprehensive coverage of relevant job content. The situational judgment test method is also relatively inexpensive, and provides complementary coverage to the enhanced multiple-choice method. The simulation method is certainly more expensive, but was very well-received by Soldiers and our Army SMEs, and we expect that economies of scale in terms of development costs would be gained over time. We were limited in our ability to pursue the idea of using training simulators to obtain high fidelity testing at lower cost, but this is a strategy that still warrants serious consideration in those few instances in which it might be workable. Moreover, it would ideally be the case that multiple measurement methods would be used for each MOS to obtain the most comprehensive and accurate assessment of technical competence. It remains the case, however, that MOS-specific testing that would involve an enhanced multiple-choice test, supplemented in some cases by a situational judgment test, would be much more affordable (and provide a reasonably comprehensive and reasonably well-accepted test experience) than testing that involves higher cost simulation or hands-on tests.

Although we did not build parallel forms of the prototype tests, this would be a requirement for most operational tests (e.g., to allow for retesting and to help maintain test security). Again, the enhanced multiple-choice test method has the advantage of well-known and accepted practices for creating multiple forms. As we discussed in Phase II, additional research is

needed to help establish such practices for situational judgment tests, but we would not expect that requirement to be a long-term issue (Knapp & Campbell, 2006).

With regard to the logistics of test administration, we experienced some issues with the Internet-delivery of the tests. Although the problems were rectified and experience with an operational system will help minimize future difficulties, technology will remain fallible. That is, in an operational program provisions will need to be made for the loss of test administration time and data due to imperfect Internet connections.

Products

Table 8.1. Summary of PerformM21-Related Products

Product	Source
Prototype automated test design survey	"Lessons Learned" Analysis (Moriarty, Knapp, & Campbell, 2006)
Common core test blueprint	"Lessons Learned" Analysis
Enhanced multiple-choice test items	
Common core test (282 items)	PerformM21 Phase I "Lessons Learned" Analysis
Armor Crewman (19K)[a]	Select21 project (Knapp, Sager, & Tremble, 2005)
Wheeled Vehicle Mechanic (63B)	PerformM21 Phase II (Knapp & Campbell, 2006)
Health Care Specialist (91W)	PerformM21 Phase II
Situational judgment test items	
Army-wide (LeadEx)	NCO21 project (Knapp et al. 2002)
Military police (31B)	PerformM21 Phase II
Health Care Specialist (91W)	PerformM21 Phase II
Simulation assessments	
EST 2000 marksmanship test	PerformM21 Phase II
Patriot Operator/Maintainer (14E)	PerformM21 Phase II
Test preparation materials (core exam and LeadEx)	
Test preparation guide	PerformM21 Phase I (R. C. Campbell, Keenan, Moriarty, & Knapp, 2004)
Self-assessment exercise	Self-assess project (Keenan & Campbell, 2005)
Soldier test score feedback reports	Phase II

[a] Although only the 19K test was used in PerformM21, the Select21 project produced comparable tests for five additional MOS (11B, 19D, 31U/25U, 74B, and 96B).

Table 8.1 lists the major products resulting from the PerformM21 research and related projects. The products vary in terms of their comprehensiveness and readiness for operational use. For example, the research (in conjunction with a companion analysis, Moriarty et al., 2006) has produced a "bank" of close to 300 common core enhanced multiple-choice test items. Along with the updated common core test blueprint developed by Moriarty et al. (2006) and the prototype test preparation materials developed based on an earlier version of the common core test and LeadEx, provide a strong starting point for an operational Army-wide assessment program. The MOS-specific products are less complete in their coverage, even for the five MOS included in the research, but would still allow a good running start on the design and development of operational tests in these five MOS.

Concluding Remarks

While we were unable to tryout all elements and strategies associated with the notional operational Army test program outlined in the PerformM21 effort, this was due primarily to limitations in resources available for the research and not to technical or logistical issues that could not be addressed in an operational program. Thus, implementation of a new Army test program appears to be feasible from this perspective.

REFERENCES

Bridge, R. G., Judd, C. M., & Moock, P. R. (1979). *The determinants of educational outcomes.* Cambridge, MA: Ballinger.

Campbell, J. P., & Knapp, D. J. (Eds.) (2001). *Exploring the limits in personnel selection and classification.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Campbell, R.C., Keenan, P. A., Moriarty, K. O., Knapp, D. J., & Heffner, T.S. (2004). *Army enlisted personnel competency assessment program phase I (Volume II): Demonstration competency assessment program development report* (Technical Report 1152). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Cullen, M. J., Sackett, P. R., & Lievens, F. (2004, April). *Threats to the Operational Use of Situational Judgment Tests In the College Admission Process.* Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, Orlando, FL.

Department of the Army. (2002). *The Army training and leader development panel report (NCO).* Final Report. Fort Leavenworth, KS: U.S. Army Combined Arms Center and Fort Leavenworth.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51*, 327-358.

Jencks, C. (1972). *Inequality.* New York: Basic Books.

Knapp, D. J., Burnfield, J. L, Sager, C. E., Waugh, G. W., Campbell, J. P., Reeve, C. L., et al. (2002). *Development of predictor and criterion measures for the NCO21 research program* (Technical Report 1128). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Knapp, D. J., & Campbell, R. C. (Eds.) (2006). *Army enlisted personnel competency assessment program: Phase II report* (Technical Report 1174). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Knapp. D.J., & Campbell, R.C. (2004). *Army enlisted personnel competency assessment program Phase I (Volume I): Needs analysis* (Technical Report 1151). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Keenan, P. A., & Campbell, R. C. (2005). *Development of a prototype self-assessment program in support of soldier competency assessment* (Study Report 2006-01). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Knapp, D. J., Sager, C. E., & Tremble, T. R. (Eds.) (2005). *Development of experimental Army enlisted personnel selection and classification tests and job performance criteria* (Technical Report 1168). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86*, 730-740.

Millsap, R. E. (1998),. The statistical analysis of method effects in multitrait-multimethod data. In P.E. Shrout & S.T. Fiske (Eds.) *Personality Research, Methods, and Theory*. Mawah, NJ: Lawrence Erlbaum Associates, Inc.

Moriarty, K. O., Knapp, D. J., & Campbell, R.C. (2006*). Incorporating lessons learned into the Army competency assessment prototype* (Study Report 2006-08). Alexandria, VA: Human Resources Research Organization.

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology, 75*, 640-647.

Office of the Under Secretary of Defense, Personnel and Readiness. (2004) *Population representation in the military services*. Retrieved May 11, 2006, from http://new.humrro.org/poprep04/appendix/appendix.html.

Peterson, N. G., Mumford, M. D., Borman, W. C., Jeanneret, P. R., & Fleishman, E. A. (Eds.) (1999). *An occupational information system for the 21st century: The development of O*NET*. Washington, DC: American Psychological Association.

Rosenthal, D., Sager, C. E., & Knapp, D. J. (2005). *A strategy to produce realistic, cost-effective measures of job performance* (Study Note 2005-03). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher eduction: Propects in a post-affirmative-action world. *American Psychologist, 56*(4), 302-318.

Waugh, G. W. (2004). Situational Judgment Test. In D. J. Knapp, R. A. McCloy, & T. S. Heffner (Eds.), *Validation of measures designed to maximize 21st-century Army NCO performance* (Technical Report 1145). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Waugh, G. W., & Russell, T. L. (2005). Criterion Situational Judgment Test (CSJT). In D. Knapp, C. Sager, & T. Tremble (Eds.), *Development of experimental Army enlisted personnel selection and classification tests and job performance criteria* (Technical Report 1168). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Whitmire, R. (2006, March 6). Concern over boys' college enrollment numbers. *NPR: Morning Edition* [Radio broadcast]. Washington, DC: National Public Radio.

APPENDIX A

MILITARY POLICE (31B) JOB ANALYSIS SURVEY[7]

Executive Summary

This document summarizes the results of a Military Police (31B) web-based job analysis survey conducted as part of *Performance Measures for the 21st Century* (PerformM21), a project being sponsored by the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) with contract support provided by the Human Resources Research Organization (HumRRO). The primary objective of the survey was to investigate how the Army's training-oriented occupational analysis process (the Occupational Data Analysis, Requirements, and Structure [ODARS] program) could be adapted to provide data for developing test specifications. Of particular interest was using the survey results to develop a prototype blueprint for a test to evaluate the competence of E4 MPs seeking promotion to E5.

The survey was created using the AUTOGEN survey development and delivery system, which serves as the basis for the data collection and analysis activities of the ODARS program. The goal of the survey was to identify the most important tasks for performance as an E4 MP. The identification of tasks to be evaluated was an iterative process that involved (a) reviewing tasks from 31B Soldier training publications (STPs), (b) combining closely related tasks into broader task statements, and (c) obtaining input on the tasks from subject matter experts (i.e., 31B Advanced NCO Course instructors and students). The final survey comprised 106 Skill Level 1 and 2 tasks. In addition, with input from the subject matter experts, we developed 18 higher-order categories to help organize the individual tasks.

The survey, administered in the fall of 2004, was sent via email to all E4-E6 Soldiers within the 31B MOS. The survey was housed on ARI's Occupational Analysis Office (OAO) occupational survey server and was available for Soldiers to complete for approximately one month. Complete responses were obtained from 386 supervisors (E5/E6) and 44 incumbents (E4).

Analysis of the survey data revealed that the tasks vary greatly in their importance to performance as an E4 MP, with the "React to mine strike/Improvised Explosive Devices" receiving the highest ratings (mean = 4.55 on a 5-point scale) and "Use hand-and-arm signals to direct traffic" receiving the lowest ratings (mean = 2.47 on a 5-point scale). Of the 18 categories of tasks, Combat Techniques (12 tasks) emerged as the most important task category, followed by Apprehend Subjects (11 tasks), Respond to Special Situations (10 tasks), Weapons (9 tasks), and MP Forms and Reports (10 tasks). A noteworthy finding was that different groups of survey respondents, such as supervisors and incumbents and respondents with and without recent deployment experience, appeared to have very similar perceptions about the relative importance of the tasks to E4 job performance. The survey results were used to design a prototype blueprint that specifies the percentage of test content (for an E4 competency assessment) that should be devoted to each task category.

[7] Prepared by Andrea Sinclair, Chad Van Iddekinge, and Deirdre Knapp of the Human Resources Research Organization (HumRRO), July 15, 2005.